


3 1761 10374375 3



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743753>

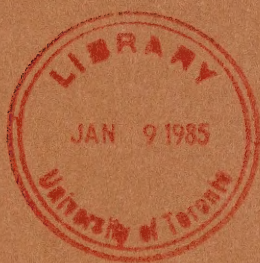
12-001



Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA



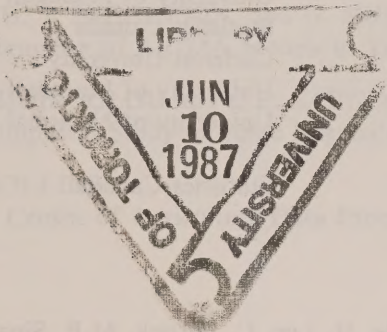
VOLUME 11, NUMBER 1
JUNE 1985

Statistics Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

June 1985



Published under the authority of
the Minister of Supply and
Services Canada

© Minister of Supply
and Services Canada 1985

December 1985
8-3200-501

Price: Canada, \$10.00, \$20.00 a year
Other Countries, \$11.50, \$23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 11, No. 1

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

EDITORIAL BOARD

Chairman	R. Platek	Statistics Canada
Editor	M.P. Singh	Statistics Canada
Associate Editors	K.G. Basavarajappa	Statistics Canada
	D.R. Bellhouse	University of Western Ontario
	E.B. Dagum	Statistics Canada
	J.F. Gentleman	Statistics Canada
	G.J.C. Hole	Statistics Canada
	T.M. Jeays	Statistics Canada
	G. Kalton	University of Michigan
	C. Patrick	Statistics Canada
Assistant Editor	J.N.K. Rao	Carleton University
	C.E. Särndal	University of Montreal
	V. Tremblay	University of Montreal
Assistant Editor	H. Lee	Statistics Canada

MANAGEMENT BOARD

R. Platek (Chairman), E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

EDITORIAL POLICY

The Survey Methodology Journal will publish articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, smoothing and extrapolation methods, demographic studies, data integration and analysis and related computer systems development and applications. The emphasis will be on the development and evaluation of specific methodologies as applied to actual data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 11, Number 1, June 1985

CONTENTS

R. PLATEK and G.B. GRAY	
Some Aspects of Nonresponse Adjustments.....	1
J.N.K. RAO	
Conditional Inference in Survey Sampling	15
G.H. CHOUDHRY, H. LEE, and J.D. DREW	
Cost-Variance Optimization for the Canadian Labour Force Survey	33
K, CHIU, J. HIGGINSON, and G. HUOT	
Performance of ARIMA Models in Time Series	51
M.A. HIDIROGLOU and C.E. SÄRNDAL	
An Empirical Study of Some Regression Estimators for Small Domains.....	65
D.K. HOLLINS	
1981 Census of Agriculture Data Processing Methodology.....	79

Some Aspects of Nonresponse Adjustments

R. PLATEK and G.B. GRAY¹

ABSTRACT

Unit and item nonresponse almost always occur in surveys and censuses. The larger its size the larger its potential effect will be on survey estimates. It is, therefore, important to cope with it at every stage where they can be affected. At varying degrees the size of nonresponse can be coped with at design, field and processing stages. The nonresponse problems have an impact on estimation formulas for various statistics as a result of imputations and weight adjustments along with survey weights in the estimates of means, totals, or other statistics. The formulas may be decomposed into components that include response errors, the effect of weight adjustment for unit nonresponse, and the effect of substitution for nonresponse. The impacts of the design, field, and processing stages on the components of the estimates are examined.

KEY WORDS: Nonresponse; Imputation; Estimation.

1. INTRODUCTION

As survey data are gathered from sampled unit, unit and item nonresponse will occur for at least some units despite all efforts to avoid it. The problem of dealing with nonresponse and the resultant missing data is two-fold. First, the effort through callbacks, repeated mailings etc. must be determined to the extent that it is cost-effective in reducing the mean square error of survey data and second, for the remaining nonresponse, the adjustments for the missing data must be obtained in order to reduce the nonresponse bias.

The field or survey centre effort to reduce or minimize unit nonresponse often means repeated attempts to contact selected units until a responsible person is available to reply to the survey questionnaire. The attempts pertain either to personal or telephone interview. In the case of mail surveys, repeated attempts mean successive mailings of a survey questionnaire to nonresponding units. In some cases, the repeated attempts may result in telephone or personal follow-ups. Some nonresponse is inevitable although every reasonable attempt should be made to minimize its levels. Thus, there will always remain some nonrespondents for whom all the efforts to convert them seem insufficient or inappropriate. The result is some imputation procedure to account for the missing data. This paper addresses the problems of controlling nonresponse at the design and field stage, followed by an examination of nonresponse adjustments at the processing stage. The examination will consider the feasibility and the practical as well as the methodological issues pertaining to the nonresponse adjustments.

Item nonresponse is often a more complex problem to deal with than unit nonresponse which is the type mostly referred to above. The most important factors which may reduce item nonresponse are good questionnaire design and a high quality of interviewers through proper hiring and training. A poorly designed questionnaire may also result in problems of following or completing the proper sequence of questions, whether by an interviewer or in a self-interview situation. Consequently, item nonresponse may occur in a questionnaire without the interviewer or respondent being aware of it. In addition, respondents may be willing to answer some but not all questions in a survey. Whatever the reason for missing items, the problems of substituting for them remains. Usually, a survey organization is unwilling to throw out whatever information

¹ R. Platek and G.B. Gray, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

has been obtained unless of course the responses to major items appear very faulty or illogical. Thus, other means of imputing for missing items while maintaining the partial information on the records are usually undertaken.

Various statistics are required from a survey or census to explain social phenomena, determine socio-economic policies, etc. These include means, totals, ratios, distributions, percentiles and graphs. The statistics are assumed to be based on a universe of N units that belong to the target population; where N may or may not be known.

It may be demonstrated that all of the statistics mentioned above may be expressed in terms of totals or counts. Consequently, the remainder of the article will deal with missing data as they affect estimates of totals and counts in surveys. Some references to censuses will also be made.

2. ESTIMATION FORMULA

In the presence of unit and item nonresponse, the estimate of the total of characteristic y may be given by the general expression as in (2.1) below.

$$\tilde{Y} = \sum_{i=1}^N t_i \pi_i^{-1} \left\{ \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}] + (1 - \delta_i) z_i \right\}, \text{ where} \quad (2.1)$$

t_i = 1 or 0 according as unit i is selected or not,

π_i = probability that unit i is selected.

δ_i = 1 or 0 according as unit i responds or not,

δ_{iy} = 1 or 0 according as responding unit i responds to item or characteristic y or not,

y_i = observed response for characteristic y when $\delta_{iy} = \delta_i = 1$;
 y_i may or may not = Y_i , the true value,

z_{iy} = imputed value for item nonresponse, when $\delta_i = 1$, $\delta_{iy} = 0$.

z_i = imputed value for unit nonresponse when $\delta_i = 0$.

The above estimate may pertain to a class a of units, when one inserts the indicators variable β_{ia} equal to 1 or 0 after π_i^{-1} to indicate whether or not unit i belongs to class a (e.g., age-sex class a).

In the case of item nonresponse, z_{iy} is nearly always an explicit *imputed value* for the missing information. The imputed value may be obtained by (i) a hot deck procedure i.e., substitution of an available response of characteristic y from the survey questionnaire of another unit that responded with respect to the characteristic and that is as similar as possible to unit i according to a decision table, (ii) substitution from other sources of data from the same unit such as an earlier survey, census, or administrative data if such data are available, (iii) by regression methods or (iv) by logical deduction and the list is by no means exhaustive. In some cases, systematic errors may occur from, for example, faulty coders or keypunchers. In such cases one attempts to change the codes to logical values relative to other information on the questionnaire in place of imputation. In any case, one hopes to achieve an imputed value or altered code as close to the true value Y_i as possible. In the case of continuous surveys, with characteristics that are stable over a long period of time (such as employment in some industries and occupations), the response or earlier survey data may be considered almost as good as that of current survey data for the same unit. This would be especially when the reference periods of the current and earlier survey data are not too far apart in time. This may be also true in the case of survey data one year apart in the case of seasonal characteristics such as, for example, those related to the fishing industry. Sometimes the imputation of earlier survey data may be used also for unit

nonrespondents that were respondents previously and with stable characteristics.

Usually, in the case of unit nonresponse, the imputation is undertaken by weight adjustment by the inverse response rate in a cell or area. The estimate of total is then given by:

$$\tilde{Y} = \sum_{i=1}^N t_i \pi_i^{-1} (wa)_i \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}] \quad (2.2)$$

where $(wa)_i$ = weight adjustment for unit i to compensate for the deficient sample due to unit nonresponse. In the above expression, it is assumed that all item nonresponse has already been imputed for by z_{iy} in the case of responding unit i when $\delta_{iy} = 0$.

The estimates of the cumulative distribution function from the sample in the context of potential missing data may be obtained by replacing the observed value y_i by the indicator variable $c(y_i, Y) = 1$ or 0 according as $y_i \leq$ or $> Y$ and similarly for z_{iy} and z_i . The estimated c.d.f.'s corresponding to (2.1) and (2.2) are respectively given by (2.3) and (2.4) below.

$$\tilde{F}(Y) = \frac{1}{\hat{N}} \sum_{i=1}^N t_i \pi_i^{-1} \{ \delta_i [\delta_{iy} c(y_i, Y) + (1 - \delta_{iy}) c(z_{iy}, Y)] + (1 - \delta_i) c(z_i, Y) \} \quad (2.3)$$

where $\hat{N} = \sum_{i=1}^N t_i \pi_i^{-1}$ denotes the estimated or the true count of units in the universe. Thus, depending upon the frame, sample design, and listings of units, \hat{N} may or may not = N .

$$\tilde{F}(Y) = \frac{1}{\hat{N}} \sum_{i=1}^N t_i \pi_i^{-1} (wa)_i \delta_i [\delta_{iy} c(y_i, Y) + (1 - \delta_{iy}) c(z_{iy}, Y)] \quad (2.4)$$

While \tilde{Y} , as defined in (2.1) and (2.2), is identical according as to whether imputation for unit nonresponse is regarded as a substitution of mean values of respondents or as a weight adjustment, the c.d.f. estimates, $\tilde{F}(Y)$ as defined in (2.3) and (2.4), are not identical. When the mean of respondents, either overall or in adjustment cells defined for compensation of nonresponse, is substituted for each missing value as in (2.1) or (2.3), there results a spiking of such mean values in the estimated c.d.f., not reflecting the real shape of the c.d.f. in the population. The use of the weight adjustment $(wa)_i$, to inflate the sample weight π_i^{-1} in (2.4) avoids this spiking effect, yielding a different but more realistic estimate of the c.d.f.

Under full unit and item response, the estimates (2.1) and (2.2) simplify to the Horvitz-Thompson (1952) estimate of the total, which is unbiased apart from response errors. In the presence of missing data and imputation for them, the estimates (2.1) and (2.2) however are likely to be biased for reasons other than response errors unless z_{iy} 's and z_i 's tend to equal y_i 's when imputation for either item or unit nonresponse is required.

In the next section, the estimates (2.1) and (2.2) are decomposed into various components due to response error, imputation error due to item nonresponse, imputation error due to unit nonresponse and the effect of weight adjustments exceeding one.

3. Components of the Estimate

The estimate \tilde{Y} given by (2.1) or (2.2) may be split up into 5 components, beginning with the Horvitz-Thompson estimate using the true values of the characteristic as in Table 1. The estimated c.d.f. $\tilde{F}(Y)$ as in (2.4) may be similarly split up but will be omitted in this paper.

When the weight adjustment $(wa)_i = 1$, the last line cancels out and the first 4 lines (3.1) to (3.4) total the estimate as given by (2.1). When the unit nonresponse is compensated for by a weight adjustment $(wa)_i > 1$, there is no direct substitution z_i for the missing value

Table 1:
Components of the Estimate \tilde{Y}

$\tilde{Y} = \sum_{i=1}^N t_i \pi_i^{-1} Y_i$..	unbiased estimate based on full response, with true values	(3.1)
$+ \sum_{i=1}^N t_i \pi_i^{-1} (y_i - Y_i)$..	effect of response error	(3.2)
$+ \sum_{i=1}^N t_i \pi_i^{-1} \delta_i (1 - \delta_{iy}) (z_{iy} - y_i)$..	effect of item nonresponse	(3.3)
$+ \sum_{i=1}^N t_i \pi_i^{-1} (1 - \delta_i) (z_i - y_i)$..	effect of unit nonresponse	(3.4)
$+ \sum_{i=1}^N t_i \pi_i^{-1} [(wa)_i - 1] \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}]$..	effect of weight adjustment for unit nonresponse	(3.5)

and z_i is taken to be 0 in (3.4). In that case, the 5 lines total the estimate as given by (2.2) and the negative effect of unit nonresponse in (3.4) is compensated for by the positive effect of weight adjustment in (3.5).

(a) Response error

The sum of the 1st and 2nd lines of the estimate \tilde{Y} (See 3.1 and 3.2) equal the desired Horvitz-Thompson estimate of total under full response. The observed response y_i for unit i may not equal the true value Y_i so that a response error at unit i level may result. The response error, which is not the real subject of this paper, can only be reduced, though not likely eliminated, by proper interviewer training, good questionnaire design with unambiguous definitions of characteristics and questions and without cluster that would confuse the interviewer and/or respondent.

When the sampled weighted response errors of (3.2) do not cancel out, the estimate of the total \tilde{Y} under full response, contains response error and upon taking expected value over all possible samples and response E_1 and E_3 (See Platek and Gray 1983), it may be found to be subject to response bias B_r and response variance in addition to sampling variance (SV). The response variance may be decomposed into simple (SRV) and correlated response variance (CRV) components.

The response bias, and all of the variance components (SV), (SRV) and (CRV) for the above estimate are derived in Platek and Gray (1983), subsection 2.2, pp. 257-8.

Response errors are usually studied by means of a reconciled reinterview program, whereby a subsample of responding units are reinterviewed and any observed differences between the original and reinterview data pertaining to the sample reference period are reconciled to determine which of the original or reinterview is the correct response. Reconciled reinterview surveys are undertaken in both the Canadian Labour Force Survey and the U.S. Current Population Surveys (CPS), two similar monthly surveys to measure unemployment, employment. etc.

For example, Poterba and Summers (1984), present in Table 2 some CPS results for a reconciled Reinterview Survey of May, 1976, based on a subsample of 3,329 men and 3,750 women. By means of reconciliation of a reinterviewed subsample, the *true* status of an individual is obtained so that it can be determined whether or not that individual responded correctly or not in the original survey, which in this case is CPS. Thus, the number of individuals with the true characteristics *Employed* in the reconciled interview sample who were actually reported as Employed, Unemployed, or Not in the LF in the original survey may be determined. From the three numbers, the proportion (or the probability) of correct and incorrect responses by true LF status may be estimated as in the table below.

Thus, for all of the men who were actually unemployed, 0.8720 is the estimated proportion of such men according to the reconciled reinterview study, who were accurately reported as unemployed while (0.0474 + 0.0806) or 0.1280 of the unemployed men were incorrectly reported as either *Employed* or *not in the Labour Force*. Thus, if y denotes characteristic *unemployed* i.e. $Y_i = 1$ when individual no. i is actually unemployed and a male then $y_i = 1$ correctly with probability 0.8720 while $y_i = 0$, incorrectly with probability 0.1280.

In the Canadian Labour Force Survey, the reconciled reinterview study sample during Jan.-Nov., 1984 covered 7,148 individuals and the corresponding probabilities of reporting labour force status as employed, unemployed or NILF in the regular LFS by *true* status as determined by the reinterview during 1984 are given in Table 3 below.

Thus the probability of correctly labelling an individual as unemployed, given that he/she actually unemployed is estimated to be .8691 in LFS compared with .8602 in CPS, almost

Table 2
Probabilities of Reporting Labour Force Status as Employed,
Unemployed, or NILF in the Regular CPS, by *True* Status as
Determined by the Reinterview Survey, May 1976.

True Status	Status as Reported in the Regular CPS		
	Employed	Unemployed	NILF
Total ¹			
Employed	0.9905	0.0016	0.0079
Unemployed	0.0356	0.8602	0.1041
NILF	0.0053	0.0025	0.9923
Men ²			
Employed	0.9922	0.0013	0.0065
Unemployed	0.0474	0.8720	0.0806
NILF	0.0062	0.0048	0.9890
Women ³			
Employed	0.9892	0.0019	0.0089
Unemployed	0.0194	0.8442	0.1363
NILF	0.0049	0.0015	0.9936

¹ Sampling size = 7,079

² Sampling size = 3,329

³ Sampling size = 3,750

Source: Tables were computed from "General Labour Force Status in the CPS Reinterview by Labour Force Status in the Original interview.
Both Sexes. Total. After Reconciliation.
May 1976, Bureau of the Census (unpublished)

Table 3
 Number of Individual and Probabilities of Reporting LF Status
 (in brackets) by *True* Characteristic. Jan.-Nov. 1984

True LF Characteristic (Reconciled reinterview)	Regular LFS			Total
	Employed	Unemployed	NILF	
Employed	4,082 (0.9831)	19 (0.0046)	51 (0.0123)	4,152
Unemployed	8 (0.0122)	571 (0.8691)	78 (0.1187)	657
NILF	28 (0.0120)	30 (0.0128)	2,281 (0.9752)	2,339
Total	4,118	620	2,410	7,148

the same. The corresponding probabilities for *Employed* and *Not in the Labour Force* in LFS are estimated during 1984 to be .9831 and .9752 compared with .9905 and .9923 for CPS, both somewhat lower in LFS. The reason for the difference cannot be determined at this stage. In any case, the response errors are likely more serious at national than at small area levels. For example, at national levels the response biases may be larger in magnitude relative to their sampling errors while a small area level estimate may be subject to response biases of about the same percent as at national level, but which may be much smaller than the sampling errors.

(b) Item Nonresponse and Imputation Error

The third line (3.3) of the estimate \tilde{Y} in Table 1 showed the deviation from the desired estimate \bar{Y} as a result of imputation for item nonresponse when the imputed value $z_{iy} \neq y_i$ and when the sampled weighted differences $(z_{iy} - y_i)$ over the sampled units with imputations for item nonresponse do not cancel out. Item nonresponse results from a respondent refusing to answer certain questions on the questionnaire may have been inadvertently left incompleated by either the respondent (in the case of self-enumeration) or by the interviewer. The second of the two causes of item nonresponse may result from similar causes as for response errors; i.e. complex questions with ambiguous definitions and/or an involved or cluttered questionnaire with a tendency for potential errors in following the proper path, depending upon replies to filter questions.

When item nonresponse does occur, an imputation strategy as described earlier may be undertaken, which almost always results in an explicit substitution. Crucial to data analysis at micro-levels is the need to obtain a value z_{iy} as close to the true value Y_i or at least as close to what would be the observed y_i , if the unit had responded to the question(s) that determine(s) characteristic y . There is unfortunately no way of knowing how close z_{iy} agrees with y_i except through re-enumeration of the unit, or a review and study of external sources or earlier survey data (which may not be available). The further danger of item nonresponse and the imputation for it may be the false sense of security to the data user who may not be aware or who may not be informed of the substituted value z_{iy} in place of a bonafide response at the micro-data level. The imputed value z_{iy} will tend to deviate in either direction from the true value Y_i to a greater extent than the potential response error y_i if that

unit responds to the characteristic. This may not always be the case. Unfortunately, it usually cannot be determined at the micro-level whether or not z_{iy} is less accurate than y_i would be. Even if the imputation error may sometimes be lower than the potential response error, it may further deteriorate the quality of the published statistics because of the presence of additional variance components.

Item nonresponse and response errors are often detected in the LFS by a monthly project Field Edit Module which analyzes questionnaires that failed edit for one or more questions. The distinction between response errors and item nonresponse however is often quite blurred in the analysis without probing into the individual questionnaires in detail. The common type of discrepancy is a miscoding of a question rather than item nonresponse per se. Many questions are split up into 5 or 6 different sub-categories and a miscoding may be interpreted as an item nonresponse for one sub-category and a response error for another sub-category pertaining to the same question. The analysis of the Field Edit Module deals with items (questions) but not sub-categories of the questions. The item discrepancy rate is thus difficult to define unambiguously. It pertains to a subset of questionnaires for which a specific question, say, No. q is relevant according to filter questions and decision tables. Let us suppose that out of a responding sample size of m questionnaires, question No. q is relevant for $m_q \leq m$ questionnaires. Then the discrepancy rate is the proportion of m_q questionnaires that failed edit, whether by item nonresponse or faulty coding. The ambiguity in the definition lies in whether the subset m_q should include those questionnaires with the question completed in error, those with the question left blank in error or merely those questionnaires with the question coded correctly or incorrectly. Notwithstanding the possible ambiguity in the definition, the item discrepancy rates for about 50 items as analysed for calendar year 1984 should indicate an upper bound to the fractional error in the estimates of statistics based on the items. A sample of item (defined in Table 4a) discrepancy rates for 1984 is given in Table 4 below.

Thus, for a straightforward item like (10) "Did the respondent do any work last week? Yes or No," the discrepancy rate is only 0.2%, much lower than even the national standard error. For more complex items like Nos. 12, 36, 41, 54 and 77 the discrepancy rate averages more than 10% with ranges 2 to 6% in either direction from the mean over the year. The discrepancies are corrected for, by hot deck procedures, use of last survey's responses (if available) or by logical deduction from other questionnaire data. Thus, in many instances an item discrepancy may be altered to a response subject to response rather than imputation error so that the discrepancy rates should be construed as an upper bound to the overall imputation error rates for the items.

(c) Unit Nonresponse and Weight Adjustment

In the case of unit nonresponse the two components of \tilde{Y} given by (3.4) and (3.5) must be studied together since unit nonresponse is generally compensated for by a weight adjustment (wa), rather than direct substitution z_i for a missing unit value. Weight adjustments are usually calculated by inverse rates in adjustment cells of which there are two basic types, balancing areas and weighting classes. Balancing areas are frequently design-dependent geographic areas such as a stratum, primary sampling unit, cluster, or a groups of strata or even the entire sample. Weighting classes are defined by post-strata (strata defined after sampling) formed on the basis of information available to both respondents and nonrespondents in the sample. The nonrespondent's information may be obtained from partial nonrespondents with some known characteristics even though the particular characteristic being estimated is not known for the partial nonrespondents. Alternatively, the information may be derived from external sources pertaining to the nonrespondents. Inverse response rates may be calculated for either balancing areas or weighting classes and used as weight adjustments to compensate for missing data in the cells.

Table 4
Average Discrepancy Rate by Item (defined in Table 4a)

Item	Average Discrepancy Rate	Range of Rates in 1984 (Min. to Max.)
10	0.2%	0.2% Every month
12	12.3%	10.4% to 14.3%
14	6.7%	5.7% to 8.4%
16	0.4%	0.3% to 0.5%
17	6.6%	2.0% to 9.9%
30	0.4%	0.3% to 0.5%
32	7.0%	3.0% to 11.6%
33	4.3%	1.8% to 6.0%
36	10.6%	8.1% to 12.7%
40	4.1%	1.5% to 6.8%
41	12.1%	6.2% to 19.7%
54	10.1%	7.9% to 12.1%
76	<0.1%	0.0% to 0.1%
77	15.0%	11.8% to 17.3%

Source: Internal report by Karen Switzer to P.D. Ghangurde March 4, 1985 "Some Findings on the Field Edit Module (FEM) Reports from 1984".

Table 4a
Definition of Items

(10)	Last week did (respondent) do any work at a job or business? Yes or No.
(12)	If yes to 11, "Did... have more than one job last week, was this a result of changing employers?" Yes or No.
(14)	What is the reason... usually works less than 30 hours per week, if actual response to (13) no. of hrs. worked 30.
(16)	Last week, how many hours was ... away from work for any reason whatsoever (holidays, vacations, illness, labour dispute, etc.) "00" should be filled in
(17)	What was the main reason for being away from work? (10 possible codes)
(30)	Last week did ... have a job or business at which he/she did not work? Yes or No.
(32)	Counting from the end of last week, in how many weeks will ... start to work at his/her new job? (Reply to Yes in (31), "Last week did ... have a job to start at a definite date in the future?")
(33)	Why was ... absent from work last week? (8 possible codes)
(36)	Identical to (14) but pertaining to <i>Unemployed</i> instead of <i>Employed</i> individuals.
(40)	Inthe past 4 weeks has ... looked for another job? Yes or No.
(41)	What has ... done in the past 4 weeks to find another job? (8 possible codes, 1 to 3 different codes in 1, 2, or 3 spaces).
(54)	What was the main reason why ... left that job? (9 possible codes) in response to yes to (50) has ... ever worked at a job or business (pert. to individuals permanently unable to work) and questions (51) to (53) dealing with date of last job and part/full time status. (54) is slipped if date of last job not too recent according to a pre-printed date in (52).
(76/77)	Class of worker and whether or not same as previous month, with respect to main job (76) and other job (77)

There are several types of weight adjustments available for inflation of the sample to compensate for unit nonresponse, the most common being the inverse response rate defined by the ratio of the sample size to the responding sample size in an adjustment cell. Thus, if the cell contains N_b units in its population and is represented by n_b selected units, where:

$n_b = \sum_{ieb} t_i$ the sample size in cell b which may or may not be a constant; depending on the definition of the cell,

$\hat{N}_b = \sum_{ieb} \pi_i t_i$, an estimate of the size of cell b in the population, usually N_b would not be known except in a census.

$m_b = \sum_{ieb} t_i \delta_i$ = no. of responding units in cell b , i.e., the responding sample size,

then, $(wa)_i = n_b/m_b$ when i lies in adjustment cell b . (3.6)

Before defining other possible weight adjustments, we will concentrate on the frequently applied inverse unweighted response rate in a cell as in (3.6). The estimate of the total defined by (2.2) with $(wa)_i = n_b/m_b$ may be rewritten as a special case of (2.1), with z_i given by:

$$z_i = \hat{T}_b / \pi_i^{-1} m_b, \quad (3.7)$$

where $\hat{T}_b = \sum_{ieb} \pi_i^{-1} t_i \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}]$, sample weighted total of responding units in cell b . In the case of equal sample weights in a cell, the imputed value z_i simplifies to the mean value of m_b respondents in the cell. By substituting z_i given by (3.7), into (2.1), it may be shown that the estimate is identical to (2.2) with $(wa)_i = (n_b/m_b)$. Thus, one may regard imputation for unit nonresponse as a substitution of $z_i = \hat{T}_b / (\pi_i^{-1} m_b)$ in (2.1) or as a weight adjustment to the sample weights by $(wa)_i = n_b/m_b$ in (2.2). In the case of the weight adjustment, one would set $z_i = 0$ in (3.4) in \tilde{Y} as split up into 5 components. Alternatively, one may employ the imputed value z_i as defined in (2.1) and in that case, one would set $(wa)_i = 1$ in (3.5) resulting in that component of $\tilde{Y} = 0$. Thus in order to consider the effect of weight adjustment $(wa)_i > 1$, both the negative component (3.4) and positive component (3.5) must be studied together; but to consider the effect of the implicit imputed value z_i , given by (3.7), one needs only to consider (3.4).

The weight adjustment (n_b/m_b) is used in LFS, where the adjustment cells are design-dependent psu's in non-self representing areas (NSR) and strata (subunits) of contiguous city blocks in self-representing areas (SR). In Table 5, the number of cells, the unweighted average of the weight adjustments and the frequency distribution of the weight adjustment in intervals 1-1.01, 1.01-1.02. ..., 1.10 and over are given by region/type of area for the survey, Jan. 1983.

The average weight adjustment of 1.0348 at Canada level is less than what one would expect with a nonresponse rate of about 5%. The reason for the apparent low average weight adjustment is that, for purposes of calculations of the inverse response rate, some unit nonrespondents with available responding data of the previous month for imputation purposes are treated like respondents. This applies to about 20 to 30% of the nonrespondents every month.

Table 5
 Number of Adjustment Cells, Average and Frequency Distribution of the
 Weight Adjustments by Region/Type of Area. January, 1983

		No. of cells in intervals of $(wa)_i$												
Region	No.	Aver.	1-	1.01-	1.02-	1.03-	1.04-	1.05-	1.06-	1.07-	1.08-	1.09-		
Type of Area	Cells	$(wa)_i$	1.01	1.02	1.03	1.04	1.05	1.06	1.07	1.08	1.09	1.10	1.10+	
Atl.	NSR	254	1.0250	143	6	22	21	13	13	9	7	8	2	10
Atl.	SR	123	1.0246	58	5	11	15	14	4	3	6	4	1	2
Que.	NSR	126	1.0550	72	2	8	10	10	6	8	6	0	1	3
Que.	SR	185	1.0265	106	0	7	8	23	11	4	5	7	3	11
Ont.	NSR	120	1.0333	58	1	10	11	11	8	4	2	2	2	11
Ont.	SR	252	1.0416	116	1	13	24	21	16	9	9	8	10	25
Pr.	NSR	328	1.0348	167	5	17	22	23	24	15	12	10	8	25
Pr.	SR	149	1.0306	40	23	23	20	13	8	7	3	5	4	3
BC	NSR	85	1.0468	38	3	7	8	8	2	5	1	1	1	11
BC.	SR	119	1.0412	46	4	7	15	10	7	7	7	3	3	10
Can.	NSR	913	1.0358	478	17	64	72	65	53	41	28	21	14	60
Can.	SR	828	1.0337	366	33	61	82	81	46	30	30	27	21	51
Canada		1,741	1.0348	844	50	125	154	146	99	71	58	48	35	111

Without a knowledge of the nonrespondents' characteristics, it cannot be determined precisely the threshold level beyond which the weight adjustment would become critical to result in an unacceptable bias along with an increase in the variance due to a smaller effective sample size. If the threshold is arbitrarily set for LFS at 1.05 (a level sometimes assumed by survey practitioners) then about 1/4 of the balancing units (441 out of 1,741) across Canada had critical weight adjustments of 1.05 or more in Jan. 1983. In many other surveys such as those dealing with income and expenditure, the nonresponse rate is higher overall and would likely be critical in nearly all cells if the same threshold of 1.05 is assumed.

There are other types of weight adjustments in cells. For example, one could exclude from cell b as defined above, those units that contain item nonresponse for at least one question. Let us suppose there are m_{bQ} units in cell b free of item nonresponse for the whole set of questions on the questionnaire. For $(m_b - m_{bQ})$ responding units in the cell with some item nonresponse the weight $(wa)_i = 1$, and for the remaining m_{bQ} responding units, free of item nonresponse, the weight adjustment is given by:

$$(wa)_i = [n_b - (m_b - m_{bQ})] / m_{bQ}, \text{ which exceeds } n_b / m_b. \quad (3.7a)$$

The following is the justification for applying no weight adjustment i.e., $(wa)_i = 1$, for those units in the cell with some item nonresponse but a larger weight adjustment (3.7a) than (n_b / m_b) , for those units free of item nonresponse. Records with item nonresponse likely contain response and imputation errors while records free of item nonresponse contain only response errors and with the large weight applied to records free of item nonresponse, it may be possible to obtain estimates with lower mean square error than by using the same

weight adjustment for all m_b responding units in the cell. To our knowledge, weight adjustments such as described above have not been applied but they may be worthy of study if the decrease in the bias offsets the increase in the variance that would occur with the different weights.

In the case of units with unequal probability sampling, there exists a weight adjustment based on the weighted sample and responding units in a cell instead of the unweighted ones. In such as case,

$$(wa)_i = \hat{N}_b / \hat{M}_b, \quad (3.8)$$

where $\hat{M}_b = \sum_{i \in b} \pi_i^{-1} t_i \delta_i$ is the sample weighted count of responding units in cell b . For the analogous case to the weight adjustment $(wa)_i$ in (3.7a) applied only to responding units free of item nonresponse,

$$(wa)_i = [\hat{N}_b - (\hat{M}_b - \hat{M}_{bQ})] / \hat{M}_{bQ} \quad (3.9)$$

where $\hat{M}_{bQ} = \sum_{i \in b} \pi_i^{-1} t_i \delta_i \Pi_{q=1}^Q \delta_{iq}$, the weighted count of responding units in cell b , free of item nonresponse.

$\delta_{iq} = 1$ or 0 according as unit i responded or did not respond to question no. q of the survey questionnaire containing Q questions; thus, $\Pi_{q=1}^Q \delta_{iq} = 1$ only if responding unit i is free of item nonresponse.

The justification for using (3.9) in lieu of (3.8) may be similar to that for using (3.7a) instead of (3.6). The justification for using weighted in place of unweighted response rates needs explanation and is provided after Table (6).

One could derive separate $(wa)_i$ expressions as of (3.7a) or (3.9) for each question q or for each characteristic y , defined by a set of one or more questions. Unfortunately, one would be faced with different weight adjustments in an adjustment cell for different questions or characteristics resulting in inconsistencies among different characteristics in published tables. In order to ensure uniform survey weights and weight adjustments, $(wa)_i$ should depend only on the unit and not on the question or characteristic though one may permit imputations for some items while excluding them for other items such as major ones in the weight adjustments (3.7a) or (3.9) as long as the inclusions and exclusions are consistent in the adjustment cell. For example, one may consider an imputation for missing item by logical deduction rather than by hot decking as pertaining to a record free of item nonresponse for weight adjustment purposes.

For each of the above weight adjustments as in (3.6) to (3.9), it can be shown that (2.2) is a particular case of (2.1) with z_i given by a weighted or unweighted mean of respondents. Thus, the implicit imputed value z_i for nonresponding unit i for each of the four cases of weight adjustments cited above is given by the expressions in Table (6). Additional notation is required for the expressions as given below:

$$\hat{T}_b = \sum_{i=1}^{N_b} t_i \pi_i^{-1} \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}] = \text{sample weighted total of unit respondents} \quad (3.10)$$

including imputations for item nonresponse
but excluding weight adjustments by inverse
unit response rate.

$$\hat{T}_{by} = \sum_{i=1}^{N_b} t_i \pi_i^{-1} \delta_i \delta_{iy} y_i = \text{sample weighted total of unit and item respondents with respect to characteristic } y, \quad (3.11)$$

$$\hat{T}_{bQy} = \sum_{i=1}^{N_b} t_i \pi_i^{-1} \delta_i \prod_{q=1}^Q \delta_{iq} y_i = \text{sample weighted total of unit and item respondents with respect to characteristic } y, \text{ but excluding those records in the cell with imputation for any item nonresponse} \quad (3.12)$$

$$\text{Thus, } \hat{T}_{bQy} \leq \hat{T}_{by} \leq \hat{T}_b.$$

The weight adjustment $(n_b - m_b + m_{bQ})/m_{bQ} = 1 + (n_b - m_b)/m_{bQ}$ of (c) \geq the weight adjustment of (n_b/m_b) of (a) since $m_{bQ} \leq m_b$ (see Table 6). Hence, for a given response rate m_b/n_b in a cell, one may anticipate a larger variance of an estimate using (c) than one using (a). The larger variance may or may not counteract a potentially smaller imputation bias in the overall mean square error. The same holds true in the case of applying weighted response rates $(\hat{N}_b - \hat{M}_b + \hat{M}_{bQ})/\hat{M}_{bQ}$ in (d) as opposed to \hat{N}_b/\hat{M}_b in (b) since $\hat{M}_{bQ} \leq \hat{M}_b$. When pps sampling is applied, the use of weighted vs. unweighted response rates leads to another interesting result. It is shown in Platek and Gray (1983), p. 264-265 that, when the response and selection probabilities, i.e., α_i and π_i , are positively correlated, the weight adjustments with weighted response rates will tend to be higher than those with unweighted rates. Thus under the condition of positive correlation between α_i and π_i , $E(\hat{N}_b/\hat{M}_b) > E(n_b/m_b)$ and similarly, $E[(\hat{N}_b - \hat{M}_b + \hat{M}_{bQ})/\hat{M}_{bQ}] > E[(n_b - m_b + m_{bQ})/m_{bQ}]$, where $E = E_1 E_2$, the expected value overall possible samples of units and subsamples of responding units as described by Platek and Gray (1983), p. 251.

Table 6
Implicit Imputed Value for Unit Nonrespondent by
Weight Adjustment (Cell Level)

	Weight Adjustment	Reference in text	Implicit Imputed value when $i=0$	Description
(a)	n_b/m_b	(3.6)	$\hat{T}_b/(\pi_i^{-1} m_b)$	Unweighted unit response rate
(b)	\hat{N}_b/\hat{M}_b	(3.8)	\hat{T}_b/\hat{M}_b	Weighted unit response rates
(c)	$\frac{n_b - m_b + m_{bQ}}{m_{bQ}}$	(3.7a)	$\hat{T}_{bQy}/\pi_i^{-1} m_{bQ}$	Unweighted unit response rates among units free of item nonresponse
(d)	$\frac{\hat{N}_b - \hat{M}_b + \hat{M}_{bQ}}{\hat{M}_{bQ}}$	(3.9)	$\hat{T}_{bQy}/\hat{M}_{bQ}$	Weighted unit response rates among units free of items nonresponse

Note: In the case of self-weighting sample (srswor as a particular case), the implicit imputed value z_i becomes the simple mean of respondents for both cases (a) and (b), and the simple mean of respondents (excluding those with some item nonresponse) in the cases of (c) and (d).

* See appendix I for derivation.

Whatever the weight adjustment used to compensate for unit nonresponse, it is doubtful that the individual values z_i *implicit imputed* would be close to the individual true values Y_i or even to the potential observed responses y_i . The best that can be achieved with the weight adjustment is to hope that adjustment cells formed to compensate for missing data due to unit nonresponse will ensure minimum differences between the characteristics of respondents and nonrespondents in the cells. Thus, the formation and delineation of adjustment cell is most crucial for compensation regardless of the type of weight adjustment that is applied.

7. FINAL REMARKS

As seen in the sections above, there is no ready-made solution to the missing data, whatever the types that occur. The initial strategy is to minimize the occurrence of missing data to the extent possible, without incurring great cost or sacrificing the timeliness of the survey data. Every attempt should be made at the onset to prepare for some nonresponse and set up imputation strategies. If missing data occur in about the manner anticipated, then the survey data processing ought to proceed on schedule, with the appropriate substitutions or weight adjustments. Clearly, the scheduling of survey data collection, publishing, etc. can proceed in a more orderly fashion in continuous or repeated surveys than in ad hoc one-time surveys for which the survey designer may not realize, until after the fact, all the things that can go wrong such as unexpected refusals or lack of interest on the part of both interviewers and respondents.

In order to deal with the nonresponse problems it is essential to maintain a continuous study of nonresponse rates by the survey characteristic (in the case of item nonresponse), reason for nonresponse, and if possible, to extend the study to an analysis of item and unit response probabilities so that imputation biases may be estimated from the survey itself. Alternatively, model-based estimates may continue to be explored to examine the imputation bias and, furthermore, to strengthen the estimates by employing additional information.

APPENDIX

Derivation of Implicit Value z_i for Unit Nonresponse imputation

In the case of (c) and (d) of Table 6, the estimate of cell b level is given by:

$$\begin{aligned} \tilde{Y}_b &= \hat{T}_{bQy} (wa)_i + (\hat{T}_b - \hat{T}_{bQy}) \\ &= \hat{T}_b + [(wa)_i - 1] \hat{T}_{bQy} \end{aligned} \tag{A.1}$$

In case (c), $(wa)_i - 1 = (n_b - m_b)/m_{bQ}$

$$= \sum_i t_i (1 - \delta_i) / m_{bQ}$$

$$\text{or } \tilde{Y}_b = \hat{T}_b + \sum_i t_i \pi_i^{-1} (1 - \delta_i) \hat{T}_{bQy} / \pi_i^{-1} m_{bQ} \tag{A.2}$$

or by equating (A.2) to (A.1), noting the definitions of \hat{T}_b in (3.10) and \tilde{Y} in (2.1), one may see that the imputed value z_i is given by $\hat{T}_{bQy} / \pi_i^{-1} m_{bQ}$ as stated in (c) of Table (6).

Similarly, when weighted response rates are employed, the implicit imputed value z_i may be found to be $\hat{T}_{bQy} / \hat{M}_{bQ}$ as in (d) of Table (6). The results for (a) and (b) of Table (6) follow by setting $m_{bQ} = m_b$ and $\hat{M}_{bQ} = \hat{M}_b$.

REFERENCES

- HORVITZ, D.G. and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- LESSLER, J.T. (1979). An expanded survey error model. In *Incomplete Data in Sample Surveys*, Volume 3 - Proceedings of the Symposium (eds. W.G. Madow, I. Olkin, and B.D. Rubin), San Diego: Academic Press, 259-270.
- PLATEK, R. (1977). Some factors affecting nonresponse. *Survey Methodology*, 3, 191-214.
- PLATEK, R. (1980). Causes of incomplete data, adjustments and effects. *Survey Methodology*, 6, 93-132.
- PLATEK, R., and GRAY, G.B. (1978). Nonresponse and imputation. *Survey Methodology*, 4, 144-177.
- PLATEK, R., and GRAY, G.B. (1979). Methodology and application of adjustments for nonresponse. Presented at the 42nd Session of International Statistical Institute, Manila, Philippines.
- PLATEK R., and GRAY, G.B. (1983). Part V - Imputation Methodology: Total Survey Error. In *Incomplete Data in Sample Surveys*, Volume 2 - Theory and Bibliographies (eds. W.G. Madow, I. Olkin, and D.B. Rubin), San Diego: Academic Press, 249-333.
- POTERBA, J.M., and SUMMERS, L.H. (1984). Response variation in the CPS: Caveats for the unemployment analyst. *Monthly Labour Review*, March 1984. Research Summaries, 37-43.

Conditional Inference in Survey Sampling

J.N.K. RAO¹

ABSTRACT

Conventional methods of inference in survey sampling are critically examined. The need for conditioning the inference on recognizable subsets of the population is emphasized. A number of real examples involving random sample sizes are presented to illustrate inferences conditional on the realized sample configuration and associated difficulties. The examples include the following: estimation of (a) population mean under simple random sampling; (b) population mean in the presence of outliers; (c) domain total and domain mean; (d) population mean with two-way stratification; (e) population mean in the presence of non-responses; (f) population mean under general designs. The conditional bias and the conditional variance of estimators of a population mean (or a domain mean or total), and the associated confidence intervals, are examined.

KEY WORDS: Conditional inference; Conditional bias; Conditional variance; Population mean; Random sample sizes

1. INTRODUCTION

In the conventional set-up for inference in survey sampling the sample design defines the sample space S (set of possible samples s) and the associated probabilities of selection, $p(s)$. The choice of an estimator is based on the criterion of consistency or unbiasedness and on the comparison of mean square errors (MSE), under repeated sampling with probabilities $p(s)$, using the sample space S as the reference set. Thus, an estimator \hat{Y} of a population mean \bar{Y} is unbiased if $E(\hat{Y}) = \sum_{s \in S} p(s) \hat{Y}_s = \bar{Y}$, where \hat{Y}_s is the value of \hat{Y} for the sample s . The MSE of the estimator \hat{Y} is given by $MSE(\hat{Y}) = \sum_{s \in S} p(s) (\hat{Y}_s - \bar{Y})^2$, and \hat{Y} is consistent if its MSE approaches zero as the sample size increases. A consistent or unbiased estimator of $MSE(\hat{Y})$, denoted as $mse(\hat{Y})$, provides a measure of uncertainty in \hat{Y} . If \hat{Y} is unbiased or consistent, then the observed values \hat{Y}_s and $mse(\hat{Y}_s)$ provide a large sample, $(1 - \alpha)$ -level, confidence interval given by

$$I_s = \hat{Y}_s \pm z_{\alpha/2} \sqrt{mse(\hat{Y}_s)}, \quad (1)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ -point of a $N(0, 1)$ variable. The interpretation of (1) is that in repeated sampling with S as the reference set, approximately 100 $(1 - \alpha)\%$ of the intervals, I_s , will contain the true value \bar{Y} .

The comparison of unconditional mean square errors, $MSE(\hat{Y})$, is appropriate at the design stage, but the sample space S may not be the relevant reference set for inference after the sample s has been drawn, if the sample contains "recognizable subsets". The concept of recognizable subsets will be illustrated in subsequent sections through examples involving random sample sizes. The choice of relevant reference set, however, is not unique. In fact, the surveyed sample s can be viewed as unique in a real sense, but then no inference under a repeated sampling set-up can be made since the relevant reference set would contain a singleton (Holt and Smith 1979).

¹ J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

Conditional inference has attracted considerable attention and controversy in classical statistics since Fisher (1925). For instance, in testing for independence in a 2×2 table of counts, Fisher argued that the inference should be conditional on the observed row and column marginal totals even if the margins are not fixed by the design. Yates (1984) revived this problem. The choice of relevant reference set is not always clear-cut, but the following guidelines look reasonable: (1) A conditional procedure should be chosen *before* observing the data, especially in the public domain. (2) A conditioning partition of S should be chosen in such a way that the partition contains no (or little) information on the parameters of interest, i.e. the statistic indexing the partition should be an ancillary statistic (Cox and Hinkley 1974, p. 38). (3) If the sample sizes are random (e.g., domain sample sizes) and their population distribution is completely known (or at least partially known), then the inferences should be conditional on the observed sample sizes. In this context, Durbin (1969, p. 643) says “If the sample size is determined by a random mechanism and one happens to get a large sample one knows perfectly well that the quantities of interest are measured more accurately than they would have been if the sample size had happened to be small. It seems self-evident that one should use the information available on sample size in the interpretation of the result. To average over variations in sample size which might have occurred but did not occur, when in fact the sample size is exactly known, seems quite wrong from the standpoint of the analysis of the data actually observed”.

The discussion throughout the paper will be confined to conditional inference in the presence of random sample sizes, as in guideline (3) above. Even with this restriction, it will be shown that conditional inferences are not always easy to implement in practice. We begin our discussion with simple examples and then extend it to more complex problems. In the context of sample surveys, Holt and Smith (1979) provide the most compelling arguments in favour of conditional inference, although their discussion was restricted to poststratification of a simple random sample (SRS); see Section 3.1.

Lahiri (1969) pointed out the “difficulties of conveying convincingly the real import of the sample survey estimates to intelligent but lay users of statistical data”; in particular, “the fallacy in implicitly using the (sampling) standard error as a measure of precision of the *observed* (sample) estimate, illustrating this point with a number of examples drawn from the current theory”.

2. SIMPLE RANDOM SAMPLING WITH REPLACEMENT

Simple random sampling (SRS) with replacement is seldom used in practice, but it provides a simple introduction to conditional inference.

Suppose a simple random sample, s , of size n is selected from a population of size N with replacement so that S contains N^n samples s . Let ν denote the number of distinct units in s . Then ν is a random variable with possible values $1, \dots, n$. Let t_i denote the number of times the i -th population unit is included in s . Then two well-known estimators of the population mean \bar{Y} are given by

$$\bar{y}_n = \frac{1}{n} \sum_{i \in s} t_i y_i, \quad (2.1)$$

the sample mean based on all the n draws, and

$$\bar{y}_\nu = \frac{1}{\nu} \sum_{i \in s} y_i, \quad (2.2)$$

the mean based on the distinct units in s . Both \bar{y}_n and \bar{y}_ν are unconditionally unbiased under the reference set S , and the unconditional variance of \bar{y}_ν is always smaller than that of \bar{y}_n . Hence, from efficiency considerations \bar{y}_ν should be preferred over \bar{y}_n . The Horvitz-Thompson estimator

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i} = \frac{\nu}{E(\nu)} \bar{y}_\nu \tag{2.3}$$

is also unconditionally unbiased, where π_i is the probability that unit i is included at least once in the sample:

$$\pi_i = \frac{E(\nu)}{N} = 1 - \left(1 - \frac{1}{N}\right)^n.$$

The comparison of variances of \bar{y}_ν and \bar{y}_{HT} shows that \bar{y}_ν is not always better than \bar{y}_{HT} . Following Durbin's (1969) argument, it is clear that for the purpose of inference one should condition on the observed value of ν , i.e., the relevant reference set is the set S_ν of $\binom{N}{\nu}$ samples of effective size ν , and not S . Fortunately, it is easy to implement conditional inference in this case since $P(s_\nu | \nu) = \binom{N}{\nu}^{-1}$, i.e. conditionally, the observed sample, s_ν , of distinct units is a simple random sample of size ν drawn without replacement. It follows that \bar{y}_ν is conditionally unbiased, i.e. $E_2(\bar{y}_\nu) = \bar{Y}$ where E_2 denotes conditional expectation, whereas $E_2(\bar{y}_{HT}) = [\nu/E(\nu)] \bar{Y} \neq \bar{Y}$ so that \bar{y}_{HT} is conditionally biased. Hence, \bar{y}_ν should be preferred over \bar{y}_{HT} , despite the inconclusive comparison of unconditional variances. Note that \bar{y}_{HT} would be a serious underestimate if the observed ν is much smaller than $E(\nu)$. A relevant measure of uncertainty is the conditional variance, $V_2(\bar{y}_\nu)$, which is estimated unbiasedly by

$$v(\bar{y}_\nu) = \left(\frac{1}{\nu} - \frac{1}{N}\right) s_{\nu y}^2, \tag{2.4}$$

where $(\nu - 1)s_{\nu y}^2 = \sum_{i \in s} (y_i - \bar{y}_\nu)^2$ and V_2 denotes the conditional variance. The appropriate confidence interval for \bar{Y} is given by

$$I_\nu = \bar{y}_\nu \pm z_{\alpha/2} \sqrt{v(\bar{y}_\nu)}. \tag{2.5}$$

Conditionally, the confidence level of I_ν is $1 - \alpha$ approximately if ν is not small. Another variance estimator

$$v^*(\bar{y}_\nu) = \left[E\left(\frac{1}{\nu}\right) - \frac{1}{N}\right] s_{\nu y}^2 \tag{2.6}$$

is conditionally biased, although unbiased when averaged over the whole sample space, S . It follows from (2.4) and (2.6) that $v(\bar{y}_\nu) < v^*(\bar{y}_\nu)$ if $1/\nu < E(1/\nu)$ and vice versa if $1/\nu > E(1/\nu)$. Thus, the confidence interval based on (2.6) would be too narrow if $E(1/\nu) < 1/\nu$ and hence yield a confidence level less than $1 - \alpha$, and too wide if $E(1/\nu) > 1/\nu$ leading to a confidence level greater than $1 - \alpha$. It may be noted that confidence intervals that are conditionally correct are automatically correct in the unconditional framework.

3. SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT

Suppose a simple random sample of fixed size n is drawn without replacement. In the absence of recognizable subsets, the relevant reference set is the set S of $\binom{N}{n}$ samples s , each of size n , and the sample mean \bar{y}_n is unbiased and its variance is estimated unbiasedly by

$$v(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N} \right) s_{ny}^2 \quad (3.1)$$

where $(n-1)s_{ny}^2 = \sum_{i \in s} (y_i - \bar{y}_n)^2$. The resulting confidence interval is given by $I_s: \bar{y}_n \pm z_{\alpha/2} \sqrt{v(\bar{y}_n)}$ with confidence level $1 - \alpha$ approximately if n is not small.

Suppose now that recognizable subsets exist in the sense that we observe the sample configuration $\underline{n} = (n_1, \dots, n_k)$ belonging to k post-strata with known weights $W_i = N_i/N$. Ideally, stratified sampling should have been used but the strata frames were not available. The relevant reference set now is the set $S_{\underline{n}}$ of $\prod \binom{N_i}{n_i}$ samples having the realized configuration \underline{n} since the distribution of \underline{n} is completely known.

3.1 All $n_i \geq 1$

If *all* the observed $n_i \geq 1$, then the customary post-stratified estimator

$$\bar{y}_{pst} = \sum W_i \bar{y}_i \quad (3.2)$$

is conditionally unbiased given \underline{n} since $P(s|\underline{n}) = \prod \binom{N_i}{n_i}^{-1}$, i.e., conditionally the observed sample s is a stratified random sample (s_1, \dots, s_k) with strata sample sizes n_i . Here \bar{y}_i denotes the sample mean in the i -th stratum. A relevant measure of uncertainty is the conditional variance, $V_2(\bar{y}_{pst})$, which is estimated unbiasedly by

$$v(\bar{y}_{pst}) = \sum W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) s_{iy}^2 \quad (3.3)$$

provided *all* $n_i \geq 2$, where $(n_i - 1)s_{iy}^2 = \sum_{j \in s_i} (y_{ij} - \bar{y}_i)^2$ (Holt and Smith 1979). The resulting confidence interval, $I_{pst}: \bar{y}_{pst} \pm z_{\alpha/2} \sqrt{v(\bar{y}_{pst})}$, is conditionally correct. Another variance estimator

$$\begin{aligned} v^*(\bar{y}_{pst}) &= \sum W_i^2 \left[E \left(\frac{1}{n_i} \right) - \frac{1}{N_i} \right] s_{iy}^2 \\ &\doteq \left(\frac{1}{n} - \frac{1}{N} \right) \sum W_i s_{iy}^2 \end{aligned} \quad (3.4)$$

is conditionally biased, although unbiased when averaged over the whole sample space, S (assuming that $P(n_i \leq 1)$ is negligible). The conditional performance of confidence interval based on (3.4) evidently depends on the extent of divergence of the observed values $1/n_i$ from their expectations $E(1/n_i)$. It may be noted that the interval I_{pst} is also correct in the unconditional framework, provided $P(n_i \leq 1)$ is negligible for all i .

If $n_i = 1$ for some i , no conditionally unbiased variance estimator can be obtained, but it might be satisfactory to use a collapsed strata method or use the model-based solution of Hartley *et al.* (1969) originally proposed for variance estimation in stratified random sampling with one unit per stratum. Empirical studies might throw some light on the applicability of the latter methods.

The customary justification for preferring \bar{y}_{pst} over \bar{y} is that the unconditional variance of \bar{y}_{pst} is approximately equal to the variance under proportional allocation and hence smaller than the unconditional variance of \bar{y} . We are also reminded that gains in efficiency under proportional allocation are likely to be modest. It is more important, however, to note that the sample mean \bar{y} is conditionally biased:

$$E_2(\bar{y}) = \sum w_i \bar{Y}_i \neq \sum W_i \bar{Y}_i = \bar{Y}, \quad w_i = \frac{n_i}{n}, \quad (3.5)$$

and hence the resulting inferences could be conditionally incorrect.

Example 1. Suppose $k = 2$ (say, male, female strata with known projected census weights W_1 and $W_2 = 1 - W_1$, or small and big hospitals (Royall 1970)). Royall used a super-population model

$$E_m(y_i) = \beta x_i, \quad i = 1, \dots, N, \quad \beta > 0, \quad x_i > 0 \quad (3.6)$$

to demonstrate that \bar{y} is model-biased conditionally, where E_m denotes the model expectation, i.e.,

$$E_m(\bar{y}) = \beta \bar{x} \neq E_m(\bar{Y}) = \beta \bar{X} \quad (3.7)$$

unless the sample mean \bar{x} coincides with the population mean \bar{X} . In his example, x_i = number of beds in the i -th hospital, y_i = number of occupied beds in the i -th hospital, and x_1, \dots, x_N are known. Royall argues that \bar{y} leads to serious underestimation if the observed sample contains all (or mostly) small hospitals since $B_m(\bar{y}) = E_m(\bar{y}) - E_m(\bar{Y}) = \beta(\bar{x} - \bar{X})$ and $\bar{x} \ll \bar{X}$. This point can also be illustrated in our conditional framework without assuming a model. The ratio of the conditional bias of \bar{y} to the population of large hospitals, \bar{Y}_2 , may be expressed as

$$\frac{B_2(\bar{y})}{\bar{Y}_2} = (W_1 - w_1)\delta = (w_2 - W_2)\delta, \quad (3.8)$$

where $B_2(\bar{y}) = E_2(\bar{y}) - \bar{Y}$ denotes the conditional bias of \bar{y} , $\delta = (\bar{Y}_2 - \bar{Y}_1)/\bar{Y}_2$ and $0 < \delta < 1$ since the population mean, \bar{Y}_1 , of small hospitals is smaller than \bar{Y}_2 . If $w_1 = 1$ (i.e., all small hospitals observed in the sample), then $E_2(\bar{y}) = \bar{Y}_1 \ll \bar{Y}$ and hence \bar{y} is a serious underestimate. Similarly, if $w_1 \gg W_1$ (i.e., mostly small hospitals observed), then it follows from (3.8) that \bar{y} would lead to serious underestimation.

In this example, one should use the post-stratified estimator $\bar{y}_{pst} = W_1 \bar{y}_1 + W_2 \bar{y}_2$ which is conditionally unbiased unless $n_1 = 0$ or $n_2 = 0$. It might be preferable, in fact, to use a post-stratified ratio estimator

$$\bar{y}_{pst,r} = \frac{\bar{y}_{pst}}{\bar{x}_{pst}} \bar{X}, \quad (3.9)$$

where $\bar{x}_{pst} = W_1 \bar{x}_1 + W_2 \bar{x}_2$ and \bar{x}_i is the sample mean of x in the i -th stratum. The estimator (3.9) is approximately unbiased conditionally and more efficient than \bar{y}_{pst} if n is large.

Remark 1. In Royall's example, one should, in fact, use a more efficient design than simple random sampling since all the population x -values are known, e.g., stratified random sampling under x -stratification and, perhaps, optimal allocation based on the x -values.

Remark 2. Royall justifies the use of the customary ratio estimator $\bar{y}_r = (\bar{y}/\bar{x})\bar{X}$ under his model (3.6), but it cannot be justified in the conditional (repeated sampling) framework since \bar{y}_r is conditionally biased:

$$B_2(\bar{y}_r) \doteq \bar{X} \left[\frac{w_2 \bar{Y}_1 + w_2 \bar{Y}_2}{w_1 \bar{X}_1 + w_2 \bar{X}_2} - R \right], \quad R = \frac{\bar{Y}}{\bar{X}} \quad (3.10)$$

$$\neq 0$$

unless $\bar{y}_1/\bar{x}_1 = \bar{y}_2/\bar{x}_2 = R$. In the extreme case of $w_1 = 1$, $B_2(\bar{y}_r) = \bar{X}(R_1 - R)$ where $R_1 = \bar{Y}_1/\bar{X}_1$. Hence, $B_2(\bar{y}_r) \lesssim 0$ according as $R_1 \lesssim R$.

Remark 3. If the weight W_1 is unknown but \bar{X} is known, we cannot implement either \bar{y}_{pst} or $\bar{y}_{pst,r}$. Royall suggests the use of \bar{y}_r with inference conditional on the observed mean \bar{x} . However, the choice \bar{x} is somewhat arbitrary, and the conditional bias of \bar{y}_r could be quite large unless the model (3.6) is true, at least approximately.

If good prior information on W_1 is available, say $W_1^* \leq W_1 \leq W_1^{**}$ where W_1^* and W_1^{**} are known, then one could use the following ‘‘pseudo’’ post-stratified estimator of \bar{Y} :

$$\bar{y}_{pst}^* = \tilde{W}_1 \bar{y}_1 + \tilde{W}_2 \bar{y}_2, \quad (3.11)$$

where $\tilde{W}_1 = w_1$ if $W_1^* \leq w_1 \leq W_1^{**}$, $= W_1^*$ if $w_1 < W_1^*$, $= W_1^{**}$ if $w_1 > W_1^{**}$ and $\tilde{W}_2 = 1 - \tilde{W}_1$. The estimator \bar{y}_{pst}^* and its ratio analogue should perform better conditionally given (n_1, n_2) than \bar{y} and \bar{y}_r , although biased. Unconditionally, the MSE of \bar{y}_{pst}^* should be smaller than the MSE of \bar{y} , provided $W_1^* \leq W_1 \leq W_1^{**}$. One could also utilize a formal Bayesian approach to estimate W_1 by specifying a prior distribution on W_1 .

Example 2 (outliers). The problem of estimating a population mean \bar{Y} in the presence of outliers is similar to the hospital example above. Suppose the population is known to contain a small fraction, W_2 , of outliers (large observations) but W_2 is unknown, i.e. $W_1 \gg W_2$ and $\bar{Y}_2 \gg \bar{Y}_1$. Then, if the observed sample contains no outliers (i.e., $w_2 = 0$), we would say that \bar{y} is ‘‘far from the true value \bar{Y} ’’ (Chinnappa 1976) and yet \bar{y} is (unconditionally) unbiased. The meaning of this statement follows from the fact that $E_2(\bar{y}) = \bar{Y}_1 \ll \bar{Y}$, where E_2 is the conditional expectation as before.

On the other hand, we would say that \bar{y} is a serious overestimate if the sample contains outliers. This follows from (3.8) noting that $w_2 \gg W_2$ (since W_2 is very small). For instance, if $N_2 = 1$ then $w_2 = 1/n \gg W_2 = 1/N$. In this situation, we are told to modify the estimate \bar{y} by reducing the weight attached to outliers in the sample. One suggestion is to modify \bar{y} by reducing the weight attached to outliers from $1/n$ to $1/N$ and adjusting the weights for non-outliers such that the n weights sum to 1:

$$\bar{y}^* = \frac{N - n_2}{N} \bar{y}_1 + \frac{n_2}{N} \bar{y}_2. \quad (3.12)$$

The conditional relative bias of \bar{y}^* is given by

$$\frac{B_2(\bar{y}^*)}{\bar{Y}_2} = \left(w_2 \frac{n}{N} - W_2 \right) \delta, \quad (3.13)$$

whereas $B_2(\bar{y})/\bar{Y}_2 = (w_2 - W_2)\delta$. If $w_2 \frac{n}{N} - W_2 < 0$, then

$$\left|w_2 \frac{n}{N} - W_2\right| = W_2 - w_2 \frac{n}{N} < w_2 - W_2 \text{ if } 2W_2 < w_2\left(1 + \frac{n}{N}\right).$$

The inequality $2W_2 < w_2(1 + n/N)$ should be satisfied since $w_2 \gg W_2$. If $w_2 n/N - W_2 > 0$, then

$$\left|w_2 \frac{n}{N} - W_2\right| = w_2 \frac{n}{N} - W_2 < w_2 - W_2.$$

Hence, the estimator \bar{y}^* should have a smaller absolute value of conditional bias than \bar{y} .

The estimator \bar{y}^* is essentially obtained from the post-stratified estimator \bar{y}_{pst} by pretending that $N_2 = n_2$. A more satisfactory solution can be obtained by gathering good prior information on $W_1 (= 1 - W_2)$, say from census data, and then using the estimator \bar{y}_{pst}^* or the estimator based on a Bayes estimator of W_1 .

Hidiroglou and Srinath (1981) derived the conditional bias and conditional and unconditional MSE of \bar{y} , \bar{y}^* and some other modifications of \bar{y} , but they did not compare the conditional biases of \bar{y} and \bar{y}^* as above.

3.2 $n_i = 0$ for Some i

If the total sample size, n , is small or if too many post-strata chosen, then n_i could be zero for some i . The post-stratified estimator (3.2) in this case reduces to

$$\bar{y}_{pst} = \sum' W_i \bar{y}_i, \tag{3.14}$$

where \sum' denotes summation over strata with nonzero n_i . The estimator (3.14) is conditionally biased:

$$E_2(\bar{y}_{pst}) = \sum' W_i \bar{Y}_i \neq \sum W_i \bar{Y}_i. \tag{3.15}$$

It remains conditionally biased even under the strong assumption $\bar{Y}_i = \bar{Y}$ for all i , which incidentally shows that \bar{y}_{pst} could lead to serious underestimation. It is also unconditionally biased. One commonly used method to overcome these difficulties is to collapse similar strata to ensure that $n_i > 0$ for all i in the reduced set of strata. Fuller (1966) proposed a more efficient solution for the special case of $k = 2$ post-strata, but his framework is unconditional in the sense that the probability, P_1^* , of $n_1 = 0$ given that either $n_1 = 0$ or $n_2 = 0$, is brought into the picture. His estimator is given by

$$\begin{aligned} \bar{y}_F &= \frac{W_1}{P_1^*} \bar{y}_1 \text{ if } n_2 = 0 \\ &= \frac{W_2}{P_2^*} \bar{y}_2 \text{ if } n_1 = 0, \end{aligned} \tag{3.16}$$

where $P_2^* = 1 - P_1^*$. The estimator \bar{y}_F is conditionally unbiased given that either $n_1 = 0$ or $n_2 = 0$, but is conditionally biased given (n_1, n_2) , even in the case $\bar{Y}_1 = \bar{Y}_2 = \bar{Y}$.

An unconditionally unbiased estimator is given by

$$\bar{y}_D = \sum \frac{a_i}{E(a_i)} W_i \bar{y}_i, \tag{3.17}$$

(Doss *et al.*, 1979), where $a_i = 1$ if at least one unit from stratum i in the sample, $= 0$ otherwise, and \bar{y}_i is defined as \bar{Y}_i if $n_i = 0$ (note that $a_i \bar{y}_i = 0$ if $n_i = 0$ even though \bar{Y}_i is unknown). The estimator \bar{y}_D , however, is conditionally biased since

$$E_2(\bar{y}_D) = \sum' \frac{W_i \bar{Y}_i}{E(a_i)} \neq \sum W_i \bar{Y}_i = \bar{Y}.$$

It remains conditionally biased even if $\bar{Y}_i = \bar{Y}$ for all i .

Doss *et al.* criticized \bar{y}_D on the grounds that it is not translation-invariant (i.e., \bar{y}_D does not change to $\bar{y}_D + c$ when each y_i is changed to $y_i + c$, where c is an arbitrary constant), and hence that the variance of \bar{y}_D , when y_i is changed to $y_i + c$, can be made arbitrarily large by increasing c sufficiently. On the other hand, the ratio estimator

$$\bar{y}_{rD} = \frac{\sum \frac{a_i}{E(a_i)} W_i \bar{y}_i}{\sum \frac{a_i}{E(a_i)} W_i}, \quad (3.18)$$

proposed by Doss *et al.*, is translation-invariant. It is conditionally biased, but the conditional bias is approximately zero if $\bar{Y}_i = \bar{Y}$ for all i , unlike the conditional bias of \bar{y}_D . Another ratio estimator which is similar to \bar{y}_{rD} conditionally is given by

$$\bar{y}_{r(pst)} = \frac{\sum' W_i \bar{y}_i}{\sum' W_i}, \quad (3.19)$$

but it is inconsistent unconditionally, unlike \bar{y}_{rD} . Hence, \bar{y}_{rD} may be preferred to $\bar{y}_{r(pst)}$ or \bar{y}_D .

If concomitant information on *all* strata is available, then one could fit a model to the observed strata means \bar{y}_i and predict the population means of strata with $n_i = 0$. For example, if the population means \bar{X}_i of a concomitant variable are linearly related to the corresponding \bar{Y}_i , then the predicted value of a \bar{Y}_i is given by $\hat{\alpha} + \hat{\beta} \bar{X}_i = \bar{y}_i^*$ (say), where $\hat{\alpha}$ and $\hat{\beta}$ are the least squares estimators obtained by minimising $\sum' (\bar{y}_i - \alpha - \beta \bar{X}_i)^2$. The resulting estimator of \bar{Y} is given by

$$\bar{y}_{pst}^* = \sum' W_i \bar{y}_i + \sum'' W_i \bar{y}_i^*, \quad (3.20)$$

where \sum'' denotes summation over strata with $n_i = 0$. This estimator should have good conditional properties if the fitted model is adequate. It should be clear from this discussion that there is no simple solution if $n_i = 0$ for some of the strata.

4. TWO-WAY STRATIFICATION

Ingenious designs to improve the efficiency of estimators have been proposed in the literature. Bryant *et al.* (1960) proposed a design involving two-way stratification in which the sample sizes n_{ij} are zero for some strata (cells). Their method is supposed to permit estimation of the population mean even when the total sample size n is less than the total number of strata. Using proportional allocation for the marginal sample sizes (n_i, n_j) , they obtained a random allocation n_{ij} such that $E(n_{ij}) = (n_i n_j)/n = n W_i W_j$, where W_i and W_j are the row and column marginal totals of cell weights W_{ij} .

Bryant *et al.* proposed the estimator

$$\bar{y}_U = \frac{1}{n} \sum \sum n_{ij} G_{ij} \bar{y}_{ij}, \quad (4.1)$$

where $G_{ij} = n^2 W_{ij} / (n_i n_j)$ and y_{ij} may be taken as \bar{Y}_{ij} if $n_{ij} = 0$. The estimator \bar{y}_U is unconditionally unbiased. However, the distribution of n_{ij} is completely known (since all W_{ij} are known) and hence the relevant reference set is the set of samples having the observed configuration $\{n_{ij}\}$, i.e., one should treat the design as stratified simple random sampling for inference purposes. The estimator \bar{y}_U is conditionally biased:

$$E_2(\bar{y}_U) = \sum \sum \left(\frac{n_{ij} G_{ij}}{n}\right) \bar{Y}_{ij} \neq \sum \sum W_{ij} \bar{Y}_{ij} = \bar{Y},$$

noting that $E_2(\bar{y}_{ij}) = \bar{Y}_{ij}$ if $n_{ij} > 0$. It also has the defects of \bar{y}_D in the previous section which can be circumvented by using the ratio estimator

$$\bar{y}_r = \frac{\bar{y}_U}{\bar{a}_U} = \frac{\sum \sum n_{ij} G_{ij} \bar{y}_{ij}}{\sum \sum n_{ij} G_{ij}} \tag{4.2}$$

where $\bar{a}_U = \sum \sum n_{ij} G_{ij} / n$. \bar{y}_r is also conditionally biased, but the conditional bias is approximately zero if $\bar{Y}_{ij} = \bar{Y}$ for all (i, j) . The latter condition, however, may be unrealistic in the present context since the strata are different by design.

As in Section 3.1, it seems necessary to use a model connecting the sampled and non-sampled strata. A reasonable model, in the absence of concomitant information, is to assume that

$$y_{ijk} = \mu + \beta_j + \tau_i + \varepsilon_{ijk} \tag{4.3}$$

where y_{ijk} is the k -th observation in the (i, j) -th cell, β_j and τ_i are fixed effects and ε_{ijk} are independent errors with zero mean and common variance σ^2 . Unfortunately, the linear combination $\mu + \beta_j + \tau_i$ for nonsampled strata is not estimable from sample data and hence the corresponding \bar{Y}_{ij} cannot be predicted. This difficulty can be avoided by assuming that β_j and τ_i are random variables and then obtaining a predictor $\hat{\mu} + \hat{\beta}_j + \hat{\tau}_i$, but the random effects model may be less realistic than (4.3) in the present context.

Motivated by the above-mentioned difficulty, Bankier (1985) discussed a raking procedure in the context of independent stratified samples according to two different criteria of stratification. His estimator is approximately model-unbiased under the fixed effects model (4.3), while the usual Horvitz-Thompson estimator and its ratio extension are model-biased.

Bankier's method can be adapted to the two-way stratification problem. The raking ratio estimator of \bar{Y} is given by

$$\bar{y}(p) = \sum \sum \frac{G_{ij}(p)}{n} y_{ij} \tag{4.4}$$

where y_{ij} is the sample total in the (i, j) -th cell ($y_{ij} = 0$ if $n_{ij} = 0$) and $G_{ij}(p)$ are the values obtained in the p -th iteration of the raking procedure such that

$$\sum_j \frac{G_{ij}(p)}{n} n_{ij} \doteq W_i = \sum_j W_{ij} \tag{4.5}$$

and

$$\sum_i \frac{G_{ij}(p)}{n} n_{ij} \doteq W_{.j} = \sum_i W_{ij}.$$

The $G_{ij}(p)$ are obtained as follows: Let $G_{ij}(0) = G_{ij} > 0 \forall (i, j)$, and

$$\begin{aligned} G_{ij}(p) &= G_{ij}(p-1) \frac{W_{.i}}{\sum_j \frac{G_{ij}(p-1)}{n} n_{ij}} \quad \text{if } p \text{ is odd} \\ &= G_{ij}(p-1) \frac{W_{.j}}{\sum_i \frac{G_{ij}(p-1)}{n} n_{ij}} \quad \text{if } p \text{ is even.} \end{aligned} \quad (4.6)$$

Under the fixed effects model (4.3), we have

$$\begin{aligned} E_m[\bar{y}(p)] &\doteq \mu + \sum_i W_{.i} \tau_i + \sum_j W_{.j} \beta_j = E_m(\sum \sum W_{ij} \bar{Y}_{ij}) \\ &= E_m(\bar{Y}), \end{aligned}$$

i.e. $\bar{y}(p)$ is approximately model-unbiased. Since $E(G_{ij}(0)n_{ij}/n) = W_{ij}$ for the choice $G_{ij}(0) = G_{ij}$, these starting values should be good. However, we may encounter convergence problems with the raking process because of the many empty cells ($n_{ij} = 0$) resulting from the Bryant *et al.* design. We hope to investigate these convergence problems as well as the conditional properties of the raking ratio estimator (4.4) in a separate paper.

If the population means \bar{X}_{ij} of a concomitant variable x are known for *all* strata, then one could fit a model to the observed strata means \bar{y}_{ij} , as in Section 3.1. For example, the model $\bar{y}_{ij} = \beta \bar{x}_{ij} + b_j + t_i + \epsilon_{ij}$ with random effects b_j and t_i might be reasonable, where ϵ_{ij} is the sample mean of errors ϵ_{ijk} in the (i, j) -th cell. A predictor $\hat{\beta} \bar{x}_{ij} + \hat{b}_j + \hat{t}_i$ of \bar{Y}_{ij} for nonsampled strata may be used in conjunction with the observed means \bar{y}_{ij} to arrive at an estimator of \bar{Y} . This approach is similar to modelling for small area estimates, except that the parameter of interest here is the overall mean \bar{Y} rather than the individual cell means \bar{Y}_{ij} . We hope to investigate the conditional properties of alternative estimators of \bar{Y} in a separate paper.

5. NONRESPONSE

5.1 A Simple Model

Suppose m responses are obtained in a simple random sample of size n . Let W_1 denote the proportion in the response stratum and $\bar{Y} = W_1 \bar{Y}_1 + W_2 \bar{Y}_2$ the population mean, where \bar{Y}_1 and \bar{Y}_2 are the means of response and nonresponse strata respectively, and $W_2 = 1 - W_1$. In this situation, conditioning on the observed value of m can be questioned since the distribution of m depends on the unknown W_1 which is involved in the parameter of interest. Also, the sample mean \bar{y}_m of respondents is unconditionally biased because $E(\bar{y}_m) = \bar{Y}_2 \neq \bar{Y}$. Hence, it is necessary to assume a model for response mechanism even in the unconditional framework, unless a subsample of nonrespondents is also sampled.

A simple model assumes that the probability of response if contacted is the same for all units, say p^* , i.e., data are missing at random. Under this model, the distribution of m depends only on p^* , and hence we should condition on m if p^* is assumed known (or at least partially known or unrelated to \bar{Y}). Oh and Scheuren (1983) have shown that conditionally given m the sample s_m of respondents is like a simple random sample of size m from the *whole* population. Hence, \bar{y}_m is conditionally unbiased, and its conditional variance is unbiasedly estimated by

$$v_2(\bar{y}_m) = (m^{-1} - N^{-1})s_{my}^2, \tag{5.1}$$

where $(m - 1)s_{my}^2 = \sum_{i \in s_m} (y_i - \bar{y}_m)^2$. The resulting confidence interval $\bar{y}_m \pm z_{\alpha/2} \sqrt{v_2(\bar{y}_m)}$ is conditionally correct, at least approximately, if m is not small.

On the other hand, the Horvitz-Thompson estimator (p^* known):

$$\bar{y}_{HT} = \frac{m}{E(m)} \bar{y}_m = \sum_{i \in s_m} \frac{y_i}{np^*} \tag{5.2}$$

is conditionally biased, as in Section 2, although unbiased when averaged over the distribution of m . For general designs, the ratio estimator

$$\hat{Y}_{HT,r} = \frac{\sum_{s_m} \frac{y_i}{\pi_i p_i^*}}{\sum_{s_m} \frac{1}{\pi_i p_i^*}} \tag{5.3}$$

is often used on grounds of efficiency, where π_i is the probability of inclusion and p_i^* is the probability of response if contacted (assumed known) for the i -th unit. In the simple case of $p_i^* = p^*$ and simple random sampling, it is interesting to note that $\hat{Y}_{HT,r}$ reduces to \bar{y}_m . Hence, the ratio estimator might perform well in a conditional framework, for general designs.

5.2 A More Realistic Model

A more realistic model assumes that data are missing at random within post-strata with known weights W_i . Let n_i and m_i respectively denote the sample size and the respondent sample size in the i -th post-stratum. Then the joint distribution of (n_i, m_i) depends only on the W_i and the response probabilities within post-strata. Hence, we should condition on the observed value of (n_i, m_i) provided the post-stratum response probabilities are either known or unrelated to the parameters of interest, viz., the post-strata means. Conditionally, the observed sample is like a stratified simple random sample with fixed strata sizes m_i (Oh and Scheuren 1983). Hence, the estimator

$$\bar{y}_{pst,m} = \sum W_i \bar{y}_{mi} \tag{5.4}$$

is conditionally unbiased, and its conditional variance is unbiasedly estimated by

$$v_2(\bar{y}_{pst,m}) = \sum W_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_{miy}^2 \tag{5.5}$$

where \bar{y}_{mi} and s_{miy}^2 are the mean and variance of sample respondents in the i -th post-stratum, respectively.

If the W_i are unknown, it is a common practice to replace W_i in (5.4) by its estimate $w_i = n_i/n$. In this case, conditional inference can be questioned since the distribution of (n_i, m_i) depends on the unknown weights W_i and since W_i are involved in the parameter $\bar{Y} = \sum W_i \bar{Y}_i$. If partial information on W_i , in the form of bounds on W_i , is available, we can proceed with conditional inference as in Example 1, Remark 3, although the resulting estimator is still conditionally biased (but likely to be better than (5.4) with W_i replaced by w_i).

6. DOMAIN ESTIMATION (SRS)

6.1 Domain mean

Under simple random sampling (SRS), the usual estimator of a subpopulation (domain) mean, \bar{Y}_i , is given by the sample mean

$$\bar{y}_i = \sum_{j \in s_i} \frac{y_j}{n_i}, \quad n_i > 0 \quad (6.1)$$

where s_i is the sample falling in the domain and n_i is the corresponding size.

If the domain size, N_i , is known, then one should condition on the observed value, n_i . The estimator \bar{y}_i is conditionally unbiased if $n_i > 0$ since conditionally s_i is a SRS sample of fixed size n_i from the domain. An unbiased estimate of the conditional variance is

$$v(\bar{y}_i) = \left(\frac{1}{n_i} - \frac{1}{N_i} \right) s_{iy}^2, \quad n_i > 0 \quad (6.2)$$

and the resulting confidence interval $\bar{y}_i \pm z_{\alpha/2} \sqrt{v(\bar{y}_i)}$ is conditionally correct.

The estimator \bar{y}_i , however, is unstable for small domains (small areas) with small n_i . Also \bar{y}_i is not defined if $n_i = 0$. One solution to the latter problem, suggested in the literature, is to use a modified estimator.

$$\bar{y}'_i = \frac{a_i}{E(a_i)} \bar{y}_i, \quad n_i \geq 0 \quad (6.3)$$

where $a_i = 1$ if $n_i \geq 1$; $= 0$ if $n_i = 0$ and \bar{y}_i is taken as \bar{Y}_i if $n_i = 0$. The estimator \bar{y}'_i , however, is conditionally biased:

$$E_2(\bar{y}'_i) = \frac{a_i}{E(a_i)} \bar{Y}_i.$$

It is an underestimate if $n_i = 0$, and an overestimate if $n_i \geq 0$, although unconditionally unbiased. The extent of overestimation depends on the magnitude of $E(a_i) = P(n_i \geq 1)$. If, for example, $P(n_i \geq 1) = 0.75$, then $E_2(\bar{y}'_i) = (\frac{4}{3}) \bar{Y}_i$ if $n_i \geq 1$.

Sarndal (1984) proposed the following estimator in the context of small area estimation:

$$\bar{y}_{is} = \bar{y} + \frac{w_i}{W_i} (\bar{y}_i - \bar{y}), \quad n_i \geq 0, \quad (6.4)$$

where $\bar{y} = \sum w_i \bar{y}_i$ is the overall sample mean and $w_i = n_i/n$. The estimator is approximately unconditionally unbiased, but conditionally biased unless $w_i = W_i$:

$$B_2(\bar{y}_{is}) = \left(\frac{w_i}{W_i} - 1\right)(\bar{Y}_i - \bar{Y}'), \quad (6.5)$$

where $\bar{Y}' = \sum w_i \bar{Y}_i$. If $n_i = 0$, the estimator \bar{y}_{is} reduces to the "synthetic" estimator \bar{y} . The extent of under- (or over-) estimation of \bar{y}_{is} depends on both $w_i/W_i - 1$ and $\bar{Y}_i - \bar{Y}'$ and hence more complex to analyse than the bias of \bar{y}'_i . However, \bar{y}_{is} would have a larger absolute conditional bias* than \bar{y} if $w_i > 2W_i$ (and hence a larger conditional MSE). Also, the conditionally unbiased estimator \bar{y}_i has a smaller conditional variance than \bar{y}_{is} if $w_i > W_i$ (neglecting the variance of \bar{y} relative to that of \bar{y}_i) and hence smaller conditional MSE.

Hidiroglou and Sarndal (1985) proposed a modification of \bar{y}_{is} :

$$\bar{y}_{is}^{**} = \begin{cases} \bar{y}_i & \text{if } w_i \geq W_i \\ \bar{y}_{is}^* = \bar{y} + \left(\frac{w_i}{W_i}\right)^2 (\bar{y}_i - \bar{y}) & \text{if } w_i < W_i. \end{cases} \quad (6.6)$$

The estimator \bar{y}_{is}^{**} is conditionally unbiased if $w_i \geq W_i$, while its conditional absolute bias is smaller than that of \bar{y} if $w_i < W_i$. A motivation for \bar{y}_{is}^{**} is that the conditional variance of \bar{y}_{is}^* (or \bar{y}_{is}) is larger than that of \bar{y}_i (neglecting the variance of \bar{y} relative to that of \bar{y}_i) if $w_i > W_i$, while the conditional variance of \bar{y}_{is}^* is smaller than that of \bar{y}_{is} if $w_i < W_i$. However, the absolute conditional bias of \bar{y}_{is}^* is larger than that of \bar{y}_{is} if $w_i < W_i$. Hence, the choice between \bar{y}_{is}^* and \bar{y}_{is} in the case $w_i < W_i$ is not clear-cut and no simple recipe seems to exist.

Drew *et al.* (1982) proposed another sample size dependent estimator which depends on a parameter K_0 . In the SRS case and the choice $K_0 = 1$, their estimator reduces to

$$\bar{y}_{id} = \begin{cases} \bar{y}_i & \text{if } w_i \geq W_i \\ \bar{y}_{is} & \text{if } w_i < W_i. \end{cases} \quad (6.7)$$

As noted above, the choice between \bar{y}_{is} and \bar{y}_{is}^* in the case $w_i < W_i$ is not clear-cut. Consequently, the choice between \bar{y}_{id} and \bar{y}_{is}^{**} is also not clear-cut.

If N_i is unknown, the conditional argument may still be relevant provided N_i is unrelated to the parameter of interest \bar{Y}_i . It is also relevant when partial information on N_i is available, such as bounds on N_i .

If a concomitant variable x with known domain mean \bar{X}_i is available, the ratio estimator

$$\bar{y}_{ir} = \frac{\bar{y}_i}{\bar{x}_i} \bar{X}_i \quad (6.8)$$

*Sarndal's estimator, however, should perform better in the case of a one-way model. The estimator is obtained by pooling estimators of the form (6.4) over two or more groups.

and a regression-type estimator (Battese and Fuller 1981)

$$\bar{y}_{lr}^* = \bar{y}_i + \frac{\bar{y}}{\bar{x}} (\bar{X}_i - \bar{x}_i) \quad (6.9)$$

are both conditionally unbiased (approximately), but \bar{y}_{lr}^* is likely to be more efficient if a regression model (through the origin) with a common slope holds true, at least approximately, for the small areas. If the slopes are varying, then an empirical Bayes estimator, which is more complex, might be more relevant (Dempster *et al.* 1981).

6.2 Domain Total

If N_i is known, then an estimate of domain total $Y_i = N_i \bar{Y}_i$ is simply obtained by multiplying a chosen estimator of \bar{Y}_i by N_i . On the other hand, the usual unbiased estimator

$$\hat{Y}_i = \hat{N}_i \bar{y}_i = \frac{N}{n} \sum_{j \in s_i} Y_j, \quad n_i \geq 1 \quad (6.9)$$

is used if N_i is unknown, where $\hat{N}_i = N w_i$ is the unbiased estimator of N_i and $P(n_i = 0)$ is assumed to be negligible.

Suppose now that we have prior information, say $N_i^* \leq N_i \leq N_i^{**}$. Then the conditional argument may be relevant. The conditional bias of \hat{Y}_i is

$$B_2(\hat{Y}_i) = (\hat{N}_i - N_i) \bar{Y}_i. \quad (6.10)$$

It follows from (6.10) (assuming $\bar{Y}_i > 0$) that $B_2(\hat{Y}_i) > 0$, i.e., overestimation, if $\hat{N}_i > N_i$ and that $B_2(\hat{Y}_i)$ increases as the domain sample size n_i increases. Similarly, $B_2(\hat{Y}_i) < 0$, i.e., underestimation, if $\hat{N}_i < N_i$ and $|B_2(\hat{Y}_i)|$ increases as n_i decreases; the conditional bias is zero if $\hat{N}_i = N_i$.

Utilizing the prior information, we can modify \hat{Y}_i as

$$\hat{Y}_i^* = \begin{cases} N_i^* \bar{y}_i & \text{if } \hat{N}_i < N_i^* \\ \hat{N}_i \bar{y}_i & \text{if } N_i^* \leq \hat{N}_i \leq N_i^{**} \\ N_i^{**} \bar{y}_i & \text{if } \hat{N}_i > N_i^{**}. \end{cases} \quad (6.11)$$

The absolute conditional bias of \hat{Y}_i^* is smaller than that of \hat{Y}_i if either $\hat{N}_i < N_i^*$ or $\hat{N}_i > N_i^{**}$, while $\hat{Y}_i^* = \hat{Y}_i$ in the interval $N_i^* \leq \hat{N}_i \leq N_i^{**}$. Hence, \hat{Y}_i^* is conditionally better than the unbiased estimator \hat{Y}_i . Also the unconditional MSE of \hat{Y}_i^* is smaller than that of \hat{Y}_i , although \hat{Y}_i^* is unconditionally biased. Unfortunately, there is no simple way to improve upon \hat{Y}_i^* in the range $N_i^* \leq \hat{N}_i \leq N_i^{**}$. In any case, \hat{Y}_i^* should be preferred over \hat{Y}_i . Good supplementary information on the domain size is necessary in estimating a domain total efficiently.

7. GENERAL DESIGNS

Post-stratification adjustment is commonly employed in complex large-scale surveys, mainly to increase the efficiency of estimators, e.g., the age-sex adjustment in the Canadian Labour Force Survey (LFS). A general theory of unconditional inference is also available.

The estimator of total Y is given by

$$\hat{Y}_{pst} = \sum M_i \frac{\hat{Y}_i}{\hat{M}_i} \tag{7.1}$$

where \hat{Y}_i and \hat{M}_i are the usual unbiased domain estimators of the i -th post-stratum total Y_i and size M_i respectively. In the LFS, projected census counts are used for the M_i . The estimator \hat{Y}_{pst} reduces to $\sum N_i \bar{y}_i$ in the SRS case (see (3.2)) and we have already seen that $\sum N_i \bar{y}_i$ is conditionally unbiased in the SRS case (assuming all $n_i \geq 1$). However, for complex designs it seems difficult to investigate the conditional properties of (7.1); even the choice of reference set is not so clear-cut. To illustrate this difficulty, consider stratified SRS with $L = 2$ strata and $k = 2$ post-strata. If we condition on the observed post-strata sample sizes (n_{h1}, n_{h2}) in each stratum h , the theory is straightforward provided the post-strata sizes N_{hi} in each stratum are known. However, in practice we will run into problems with zero sample sizes n_{hi} and also the sizes N_{hi} in each stratum may not be available or the projections inaccurate, although $N_{.i} = \sum_h N_{hi} = M_i$ are available. Hence, we may prefer to condition on the observed total sample sizes $(n_{.1}, n_{.2})$, where $n_{.i} = \sum_h n_{hi}$.

The estimator \hat{Y}_{pst} in this special case of stratified SRS ($L = 2, k = 2$) reduces to

$$\hat{Y}_{pst} = N_{.1} \frac{y_{11} \frac{y_{11}}{n_{11}} + N_{21} \frac{y_{21}}{n_{21}}}{N_{.1} \frac{y_{11}}{n_{11}} + N_{21} \frac{y_{21}}{n_{21}}} + N_{.2} \frac{y_{12} \frac{y_{12}}{n_{12}} + N_{22} \frac{y_{22}}{n_{22}}}{N_{.1} \frac{y_{12}}{n_{12}} + N_{22} \frac{y_{22}}{n_{22}}} \tag{7.2}$$

where $N_h = N_{h1} + N_{h2}$ and $n_h = n_{h1} + n_{h2}$ are the strata population and sample sizes respectively, and y_{hi} are the sample totals in the (h, i) -th cell. The conditional expectation of (7.2) given $(n_{.1}, n_{.2})$ is not tractable since one has to evaluate the sum

$$E_2(\hat{Y}_{pst}) = \sum_t p(s_t | n_{.1}, n_{.2}) \hat{Y}_{pst}(t) \tag{7.3}$$

where s_t is a possible sample such that the observed sample sizes \tilde{n}_{hi} satisfy $\tilde{n}_{1i} + \tilde{n}_{2i} = n_{.i}$ ($i = 1, 2$), and $\hat{Y}_{pst}(t)$ is the value of (7.2) for the sample s_t , and $p(s_t | n_{.1}, n_{.2})$ is the conditional probability of observing s_t given $(n_{.1}, n_{.2})$:

$$p(s_t | n_{.1}, n_{.2}) = \left[\sum_{n_{11}=0}^{n_1} \binom{N_{11}}{n_{11}} \binom{N_{12}}{n_{11}-n_{11}} \binom{N_{21}}{n_{.1}-n_{11}} \binom{N_{22}}{n_{.2}-n_{.1}+n_{11}} \right]^{-1} \tag{7.4}$$

It is clear from (7.3) and (7.4), however, that $E_2(\hat{Y}_{pst}) \neq Y$ since \hat{Y}_{pst} does not depend on the cell totals N_{hi} unlike $p(s_t | n_{.1}, n_{.2})$.

Turning to variance estimation, the usual formula for general designs is given by

$$v^*(\hat{Y}_{pst}) = v(z_t^*) \tag{7.5}$$

where $v(y_t) = v(\hat{Y})$ is the usual variance estimator of the estimated total \hat{Y} , and $v(z_t^*)$ is obtained from $v(\hat{Y})$ by replacing y_t by

$$z_t^* = y_t - \sum_i \frac{\bar{Y}_i}{\bar{M}_i} a_t(i) \quad (7.6)$$

where $a_t(i) = 1$ if the t -th element belongs to the i -th post-stratum and $a_t(i) = 0$ otherwise (Williams 1962). In the SRS case, (7.5) reduces to

$$v^*(\hat{Y}_{pst}) \doteq N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \sum n_i s_{iy}^2 \quad (7.7)$$

(assuming $(n_i - 1)/(n - 1) \doteq n_i/n$) which is not equal to (3.3) when multiplied by N^2 . Hence, (7.5) does not behave well in the conditional framework, even in the SRS case. On the other hand, a new variance estimator

$$v(\hat{Y}_{pst}) = v(z_t), \quad (7.8)$$

where

$$z_t = \sum_i \frac{M_i}{\bar{M}_i} (y_t(i) - \frac{\bar{Y}_i}{\bar{M}_i} a_t(i)) \quad (7.9)$$

and $y_t(i) = y_t$ if the t -th element belongs to the i -th post-stratum and $y_t(i) = 0$ otherwise, might be preferable over $v^*(\hat{Y}_{pst})$ since in the SRS case it reduces to (3.3) when multiplied by N^2 and the finite population correction is ignored:

$$v(\hat{Y}_{pst}) = \sum_i \frac{N_i^2}{n_i} s_{iy}^2. \quad (7.10)$$

Some theory for ratio estimators under models also suggests that $v(\hat{Y}_{pst})$ might perform better conditionally than $v^*(\hat{Y}_{pst})$. In any case, there is no harm in switching to (7.8) since it is asymptotically equivalent to the customary variance estimator (7.5), unconditionally.

8. DISCUSSION

Our study clearly shows that conditional inference for complex designs involves formidable difficulties. Nevertheless, we should not use conventional procedures blindly. In those cases where conditional inference is feasible, as in the SRS case, we should certainly employ conditionally relevant methods as elaborated in Sections 2 - 6, while in the more complex cases we should at least make simple modifications to conventional methods, as in (7.8), so that they agree with known, conditionally correct results in special cases. Clearly, we need more research in this area.

ACKNOWLEDGEMENTS

This paper is based on my lectures given at a workshop on conditional inference in sample surveys. I am thankful to Mrs. Nanjamma Chinnappa for organizing the workshop at Statistics Canada. Constructive comments from colleagues at Statistics Canada and Professor D. Holt were helpful in preparing this paper.

REFERENCES

- BANKIER, M. (1985). Conditionally unbiased estimators based on any number of independent stratified samples. Memorandum, Business Survey Methods Division, Statistics Canada.
- BATTESE, G.E., and FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.
- BRYANT, E.C., HARTLEY, H.O., and JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.
- CHINNAPPA, B.N. (1976). A preliminary note on methods of dealing with unusually large units in sampling from skew populations. Unpublished Technical Report, Institution and Agriculture Survey Methods Division, Statistics Canada.
- COX, D.R., and HINKLEY, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- DEMPSTER, A.P., RUBIN, D.B., and TSUTAKAWA, R.K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association*, 76, 341-353.
- DOSS, D.C., HARTLEY, H.O., and SOMAYAJULU, G.R. (1979). An exact small sample theory for post-stratification. *Journal of Statistical Planning and Inference*, 3, 235-248.
- DREW, J.H., SINGH, M.P., and CHOUDHRY, H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 17-47.
- DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling* (Eds. N.L. Johnson and H. Smith), New York: Wiley - Interscience.
- FISHER, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd (5th Ed., 1934).
- FULLER, W.A. (1966). Estimation employing post strata. *Journal of the American Statistical Association*, 61, 1172-1183.
- HARTLEY, H.O., RAO, J.N.K., and KIEFER, G. (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association*, 64, 841-851.
- HIDIROGLOU, M.H., and SÄRNDAL, C.E. (1985). An empirical study of some regression estimators for small domains. *Survey Methodology*, 11, 65-77.
- HIDIROGLOU, M.H., and SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.
- HOLT, D., and SMITH, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Ser. A*, 142, 33-46.
- LAHIRI, D.B. (1969). On the unique sample, the surveyed one. Unpublished Technical Report, Indian Statistical Institute.
- OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*, Vol. 2, Academic Press, 142-184.
- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- SÄRNDAL, C.E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- WILLIAMS, W.H. (1962). The variance of an estimator with post-stratified weighting. *Journal of the American Statistical Association*, 57, 622-627.
- YATES, F. (1984). Tests of significance for 2×2 contingency tables. *Journal of the Royal Statistical Society, Ser. A*, 147, 426-463.

Cost-Variance Optimization for the Canadian Labour Force Survey

G.H. CHOUDHRY, H. LEE, and J.D. DREW¹

ABSTRACT

The cost-variance optimization of the design of the Canadian Labour Force Survey was carried out in two steps. First, the sample designs were optimized for each of the two major area types, the Self-Representing (SR) and the Non-Self-Representing (NSR) areas. Cost models were developed and parameters estimated from a detailed field study and by simulation, while variances were estimated using data from the Census of Population. The scope of the optimization included the allocation of sample to the two stages in the SR design, and the consideration of two alternatives to the old design in NSR areas. The second stage of optimization was the allocation of sample to SR and NSR areas.

KEY WORDS: Multi-stage designs; Sample allocation; Linear cost function; Components of variance.

1. INTRODUCTION

The Canadian Labour Force Survey (LFS) is a monthly household survey conducted by Statistics Canada to produce estimates for various labour force characteristics. It follows a stratified multi-stage rotating sample design with six rotation groups. Since its inception in 1945, the survey has undergone a sample redesign following each decennial census of population. These redesigns serve to update the sample to reflect population changes. They also provide the opportunity to introduce improved sampling and estimation methodologies, and to respond to shifts in information needs to be satisfied by the survey.

The 1981 post censal redesign effort included a research phase as outlined in an earlier paper (Singh and Drew 1981) in which all aspects of the survey design were examined in an effort to improve the cost efficiency of the survey vehicle. Highlights of the research program were presented by Singh, Drew, and Choudhry (1984). This report deals with the research aimed at cost-variance optimization of the sample design.

The two important factors in the choice of a sample design are the total cost and the reliability of the resulting estimates. The optimum solution can be obtained by minimizing either total cost or total variance when the other is fixed. Equivalently, the approach we have followed is one of minimizing the product of variance and cost for fixed sample size.

The cost-variance optimization was carried out in two steps. We first consider the optimization of the sample designs followed in each of the two major area types identified in the LFS design; i.e., the SR Areas or major cities, and NSR Areas which are the smaller urban and rural areas. The scope of the optimization includes the allocation of sample to the two stages of the SR design (Section 2), and the consideration of alternatives to the old design in NSR areas (Section 3). For NSR areas the old design is first evaluated empirically via a components of variance approach, and one stage of sampling in rural areas is identified for elimination. Subsequently the modified old design is compared to an alternative design featuring explicit rural/urban stratification from an overall cost-variance perspective. For both types of areas variances are obtained empirically using data from the 1971 and 1976 Censuses, while cost models are developed using data from a time and cost study, and by means of a simulation study.

¹ G.H. Choudhry, H. Lee, and J.D. Drew, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

In Section 4, we consider the second stage of optimization, the allocation of sample to NSR and SR areas, taking into account the design improvements identified for each type of area. Finally, Section 5 summarizes the improvements identified, and their implications on the redesigned sample.

2. SR DESIGN

The old SR design is a stratified two-stage design (Platek and Singh 1976). Each Self-Representing Unit (SRU) is stratified into a number of contiguous strata called subunits and each subunit is subdivided into clusters which are the primary sampling units (PSU's). The PSU's are selected using the random group method due to Rao, Hartley, and Cochran (1962) and at the second stage of sampling, a systematic sample of dwellings is taken in such a manner that the design becomes self-weighting. Let $1/W$ be the sampling rate in the stratum and n be the number of PSU's to be selected from the stratum. The N PSU's in the stratum are randomly partitioned into n groups so that the i -th random group contains N_i PSU's and $\sum_{i=1}^n N_i = N$. Let x_j and M_j , $j = 1, 2, \dots, N$, respectively be the size measure and dwelling count for the j -th PSU in the stratum.

Define

$$\lambda_j = \frac{x_j}{\sum_{i=1}^N x_i}$$

and

$$\delta_{ij} = 1 \text{ if } j\text{-th PSU is in } i\text{-th group} \\ = 0 \text{ otherwise.}$$

Then $\pi_i = \sum_{j=1}^N \delta_{ij} \lambda_j$ is the relative size of the i -th group. Now define W_{ij} 's as

$$W_{ij} = \delta_{ij} \left[W \frac{\lambda_j}{\pi_i} \right] \text{ or } \delta_{ij} \left[W \frac{\lambda_j}{\pi_i} + 1 \right] \quad (2.1)$$

such that $\sum_{j=1}^N W_{ij} = W$ for $i = 1, 2, \dots, n$, where $[a]$ is the greatest integer less than or equal to a . Now select one PSU from each of the n random groups independently with probability proportional to W_{ij} 's and sub-sample the selected PSU j from the i -th group at the rate $1/W_{ij}$. Then the overall sampling rate within each of the random groups is $1/W$ so that the design becomes self-weighting with a design weight equal to W . The average sample size for the stratum is given by

$$m = \frac{1}{W} \sum_{j=1}^N M_j \quad (2.2) \\ = M_0/W$$

where M_0 is the total number of dwellings in the stratum. Let M_{ij} be the number of dwellings in the selected PSU j in the i -th group, then $m_i = M_{ij}/W_{ij}$ dwellings will be selected from the i -th group. The average number of dwellings selected from the i -th group for a given random grouping is $1/W \sum_j \delta_{ij} M_j$ and the average over all possible random groupings is $m N_i/N$ since the expected value of δ_{ij} is N_i/N . If $N_i/N = 1/n$, i.e., the number of psu's in each of the random groups is the same, then the average sample per selected PSU is $m/n = d$ (say), where d will be called the average density for the stratum. Since m is fixed, the sample of m dwellings can be elected by varying n and d such that the product (nd) remains equal to

m , the total sample size for the stratum. Our objective here is to obtain d which for a fixed sample size minimizes the product of variance and cost. For the optimization we obtain the total variance via the components of variance approach and consider a linear cost function as described in the following section.

2.1 Variance Function

Suppose that we are interested in the total of a characteristic y for the subunit. Let y_{jh} be the y -value for the h -th household in PSU j where $h = 1, 2, \dots, N$, then the total $Y = \sum_{j=1}^N \sum_{h=1}^{M_j} y_{jh}$ is estimated by

$$\hat{Y} = W \sum_{i=1}^n y_i \quad (2.3)$$

where y_i is the sum of the y -values for the m_i selected households from the PSU selected from the i -th group, $i = 1, 2, \dots, n$. Ignoring the effect due to rounding involved in defining W_{ij} , the variance of \hat{Y} is given by (Rao et al. 1962)

$$\text{Var}(\hat{Y}) = A \left[\sum_{j=1}^N \frac{Y_j^2}{\lambda_j} - Y^2 \right] + \sum_{j=1}^N M_j S_j^2 \left[W - 1 - A \left(\frac{1}{\lambda_j} - 1 \right) \right]. \quad (2.4)$$

where

$$Y_j = \sum_{h=1}^{M_j} y_{jh},$$

$$S_j^2 = \frac{1}{M_j - 1} \sum_{h=1}^{M_j} \left(y_{jh} - \frac{Y_j}{M_j} \right)^2,$$

$$A = \frac{\sum_{i=1}^n N_i^2 - N}{N(N - 1)}.$$

If $N_i = N/n$, i.e., all random groups have equal number of PSU's, then

$$A = \frac{N - n}{n(N - 1)}.$$

Relative variance of \hat{Y} defined by $\text{Var}(\hat{Y})/Y^2$ will be

$$\begin{aligned} \text{Rel. Var}(\hat{Y}) &= A \left[\frac{1}{Y^2} \sum_{j=1}^N \frac{Y_j^2}{\lambda_j} - 1 \right] + \frac{1}{Y^2} \sum_{j=1}^N M_j S_j^2 \left[W - 1 - A \left(\frac{1}{\lambda_j} - 1 \right) \right] \\ &= A\mu_1 + (W - 1)\mu_2 + A\mu_2 - A\mu_3 \\ &= (W - 1)\mu_2 + A(\mu_1 + \mu_2 - \mu_3) \end{aligned} \quad (2.5)$$

where

$$\mu_1 = \frac{1}{Y^2} \sum_{j=1}^N \frac{Y_j^2}{\lambda_j} - 1$$

$$\mu_2 = \frac{1}{Y^2} \sum_{j=1}^N M_j S_j^2,$$

$$\mu_3 = \frac{1}{Y^2} \sum_j M_j \frac{S_j^2}{\lambda_j}.$$

μ_1 , μ_2 , and μ_3 are the population parameters and are fixed for a particular characteristic. Since $m = nd$ and if we assume that $N_i = N/n$ then we can write A as

$$A = \frac{1}{N-1} (N \frac{d}{m} - 1)$$

and

$$\begin{aligned} \text{Rel. Var}(\hat{Y}) &= (W-1) \mu_2 + (N \frac{d}{m} - 1) \frac{(\mu_1 + \mu_2 - \mu_3)}{(N-1)} \\ &= \alpha_0 + \alpha_1 d \end{aligned} \quad (2.6)$$

where

$$\alpha_0 = (W-1) \mu_2 - \frac{(\mu_1 + \mu_2 - \mu_3)}{(N-1)}$$

$$\alpha_1 = \frac{N}{m} \frac{(\mu_1 + \mu_2 - \mu_3)}{(N-1)}.$$

From (2.6), we observe that from reliability point of view, the value $d = 1$ (i.e., one dwelling per PSU) is optimum. But this will have impact on the cost as discussed in the next section. The values of α_0 and α_1 for unemployed for Halifax SRU were obtained from 1981 census data and these are

$$\alpha_0 = 0.019005, \quad \alpha_1 = 0.0007972.$$

Since α_1 is very small as compared to α_0 , the increase in the variance with the corresponding increase in d will be very small. Next we examine the effect on the cost due to varying the value of the average density d .

2.2 Cost Model

A simple cost model has been considered to investigate the impact on the cost as the density is varied. Due to telephone interviewing in the SR areas, personal visits are only required to a PSU during the rotation month and in cases where some households were without a telephone or did not agree to telephone interviewing.

A breakdown of the interviewing cost by telephone and personal visit is available for individual interviewers from field operations, but further breakdown of the personal visit component of the cost was required to construct the cost model. For this purpose a special time and cost study was carried out in the field for a period of six months (February-July 1982) on a random sample of interviewers. The results from the analysis of time and cost data are documented in a report by Lemaitre (1983). For the purpose of our cost model, we define the following set of parameters

c_0 = Fixed costs

c_1 = Average cost of dwelling-to-dwelling travel within the same PSU

c_2 = Average cost of PSU-to-PSU travel

γ = Number of PSU-to-PSU moves per selected PSU.

The fixed cost c_0 includes the time spent actually conducting interviews whether by telephone or in person and the travel cost from home to area and back. The fixed cost c_0 depends only on the total sample size m and not on n , the number of selected PSU's. Suppose that there are g_1 dwelling-to-dwelling moves and g_2 PSU-to-PSU moves made, then the total cost for m dwellings will be

$$T = c_0 + g_1c_1 + g_2c_2. \tag{2.7}$$

If n is increased then g_2 will also increase and g_1 will decrease and vice-versa but $(g_1 + g_2)$ should remain constant because the number of moves depends on the sample size m and the proportion of households interviewed by personal visit. Then we may write

$$g_1 + g_2 = \theta m. \tag{2.8}$$

From (2.8) we substitute g_1 in equation (2.7) and obtain

$$\begin{aligned} T &= c_0 + \theta mc_1 + g_2(c_2 - c_1) \\ &= c_0 + \theta mc_1 + n\gamma(c_2 - c_1). \end{aligned}$$

Now replacing n by m/d we have

$$T = c_0 + \theta mc_1 + \frac{m\gamma}{d} (c_2 - c_1)$$

and cost per dwelling C as a function of average density d is given by

$$C = \frac{c_0}{m} + \theta c_1 + \frac{\gamma}{d} (c_2 - c_1). \tag{2.9}$$

From Time and Cost Study the parameters c_1 and c_2 for Halifax were 0.78 and 2.51 respectively. These parameters were observed with average density equal to 5 but c_2 increases with d and c_1 decreases with d . Assuming that the average distance between the units is inversely proportional to the square root of the number of units in an area, we can replace c_1 by $c_1(5/d)^{1/2}$ and c_2 by $c_2(d/5)^{1/2}$ in our model so that the modified model becomes

$$C = \frac{c_0}{m} + \theta c_1 \left(\frac{5}{d}\right)^{1/2} + \frac{\gamma}{d} \left\{ c_2 \left(\frac{d}{5}\right)^{1/2} - c_1 \left(\frac{5}{d}\right)^{1/2} \right\}. \tag{2.10}$$

c_0/m is fixed per dwelling cost and does not depend on density and its value was 3.28 from Time and Cost Study. The parameter θ does not depend on the density either and was equal to 0.356 from Time and Cost Study. The parameter γ increases with density because the average number of visits to a PSU will increase due to higher density. We have approximated γ by

$$\frac{1}{6} + \frac{5}{6} (1 - p^d)$$

where p is the probability of telephone interview for a household in a non rotate-in PSU and the value of p was 0.85 as obtained from interviewers' data. From the cost model (2.10), the values of per dwelling cost for $d = 2, 3, \dots, 10$ are given in Table 1 along with the relative variances and the products of these two which are the values of the objective function to be minimized.

Table 1
Value of Relative Variance, Cost per Dwellings and
Objective Function for Various Densities (Unemployed)

Density	Relative Variance	Cost per Dwelling	Objective Function
2	0.0206	3.79	0.078
3	0.0214	3.79	0.081
4	0.0222	3.79	0.084
5	0.0230	3.78	0.087
6	0.0238	3.77	0.090
7	0.0246	3.76	0.092
8	0.0254	3.75	0.095
9	0.0262	3.74	0.098
10	0.0270	3.73	0.101

As expected, we observe that under the model considered here, the cost per dwelling decreases very slowly as the density increases since the fixed per dwelling cost (c_0/m) dominates in (2.10) due to telephone interviewing. From the previous section we had found that the increase in the relative variance is very small as the density increases. As a result our objective function is monotonically increasing but the loss in the cost-variance efficiency with increase in d is small. However it was decided to retain the old density of 5 for the redesigned sample on the grounds that lower density would have resulted in more selected PSU's with higher implementation and maintenance costs.

3. NSR DESIGN

3.1 NSR Design Alternatives

Design Alternative D_0 : Old NSR Design (see Figure 1)

Key features of the old NSR design (Platek and Singh 1976) were:

- i) **Stratification:** Economic Regions (ER's) whose numbers varied from 1-10 per province served as major strata. Within ER's, from 1-5 geographically contiguous strata were formed, using industry data from the 1971 Census.
- ii) **Primary Sampling Units (PSU's):** These were delineated within strata, to be geographically compact areas similar to the stratum with respect to stratification variables, and with respect to the ratio of rural to urban population. PSU populations ranged from 3,000 to 5,000. In the first stage PSU's were selected following the randomized probability proportional to size systematic (RPPSS) method of Hartley and Rao (1962). Within PSU's urban and rural parts were sampled separately.
- iii) **Within PSU Sampling: Urbans** All urban centers assigned in whole or in part to selected PSU's were included in the sample. The second stage of sampling was a sample of blocks, following the RPPSS method. The third and final stage of sampling was a systematic sample of dwellings.

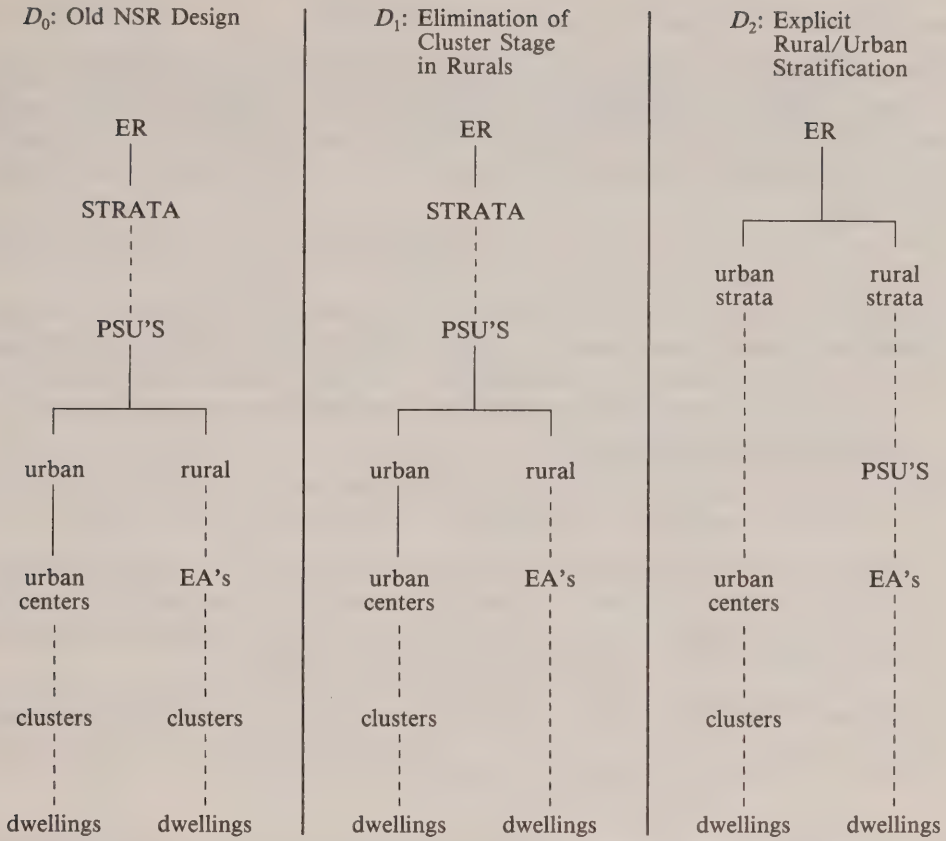


Figure 1. Representation of NSR Design Alternatives. (—— stratification, ----- stage of sampling)

iv) **Within PSU Sampling: Rurals** The second stage of sampling was a RPPSS sample of EA's. EA's were then field counted for the purposes of delineating clusters having from 3-20 dwellings. The third and fourth stages of sampling corresponded to an RPPSS sample of clusters and a systematic sample of dwellings.

Design Alternative *D*₁: Elimination of Cluster Stage of Sampling in Rurals

- i) It would permit shortening of the lead time to select independent samples from the LFS frame to 7 months from 13 months, by eliminating the need for counting of EA's.
- ii) Elimination of the clustering step would reduce sample maintenance costs.
- iii) A priori, the reduction in the stages of sampling from 4 to 3 stages would translate into a reduced variance. it was expected that costs, on the other hand, would not be very much affected, particularly with the shift to telephone interviewing.
- iv) At an early juncture in the redesign research program a field study was carried out on the operational implications of eliminating the cluster stage. Verification of EA listings a year later revealed no problems with the quality of listings, and analysis revealed no discernable impact on data collection costs.

Design Alternative D_2 : Explicit Urban/Rural Stratification

The old design with its separate sampling of urban and rural portions of PSU's featured an implicit urban/rural stratification. A drawback of the approach however was that maintenance of the stratum urban to rural population ratio at the PSU level required frequent discontinuity between rural and urban portions of PSU's, leading in turn to increased travelling costs.

In view of this problem with the old design, design alternative D_2 was formulated as follows:

- i) **Stratification:** Rural and urban portions of ER's would constitute primary strata, which would be optimally sub-stratified to the point of having strata yields of 100-150 dwellings (i.e., 2-3 PSU's each corresponding to an interviewer's assignment). ER's not able to support at least one such urban and one such rural stratum (roughly $\frac{1}{3}$ of ER's) were considered ineligible for D_2 .

Secondary rural strata would be contiguous, while secondary urban strata would be formed without geographic constraints.

- ii) **Sampling Within Rural Strata:** PSU's similar to the stratum with respect to stratification variables would be formed by grouping geographically contiguous EA's and will be selected by the RPPSS method. Second and third stages of sampling would be an RPPSS sample of EA's and systematic sample of dwellings.
- iii) **Sampling Within Urban Strata:** Sampling would proceed in three stages as follows: RPPSS sample of PSU's (individual or combined urban centers), RPPSS sample of clusters, and systematic sample of dwellings.

3.2 Variance Components Model

Design alternative D_0 , D_1 and D_2 were simulated using census data. Expressions for the variance components are given below:

Stage of Sampling	Variance Expression	
1st	$V_{(1)} = V_{(1)}^{\text{RPPSS}}$	(3.1)

2nd	$V_{(2)} = W \sum_{i=1}^N \frac{V_{(2)i}^{\text{RPPSS}}}{W_i}$	(3.2)
-----	--	-------

3rd	$V_{(3)} = W \sum_i \sum_j \frac{V_{(3)ij}^{\text{SRS}}}{W_{ij}} \text{ if last stage,}$ $= W \sum_i \sum_j \frac{V_{(3)ij}^{\text{RPPSS}}}{W_{ij}} \text{ otherwise}$	(3.3)
-----	--	-------

4th (where applicable)	$V_{(4)} = W \sum_i \sum_j \sum_k \frac{V_{(4)ijk}^{\text{SRS}}}{W_{ijk}}$	(3.4)
---------------------------	--	-------

The variance formula and its computation method for the RPPSS sampling are described in Appendix A.

3.3 Cost Model

Whereas the cost model for the SR areas dealt with allocation of samples to 2 stages of sampling, here a cost model is needed to compare alternative NSR designs.

The cost model for design D_1 under personal interviewing was formulated as

$$C_{D_1} = F_0 + F_1 + F_2 + E_1 + E_2$$

where F_0 = fixed fee for interviewing,
 F_1 = fee for home to area, between PSU, and between secondary travel,
 F_2 = fee for within secondary (dwelling to dwelling) travel,
 E_1 = expenses associated with home to area, between PSU, and between secondary travel,
 E_2 = expenses associated with dwelling to dwelling travel.

Fees are compensation for the time spent and expenses for the distance covered. All Parameters are expressed in terms of per dwelling costs.

Under telephone interviewing, this was modified to

$$C_{D_1}^T = F_0 + \alpha(F_1 + F_2 + E_1 + E_2),$$

where α is the factor by which time and mileage would be decreased under telephoning.

Now, under the assumption that D_2 would affect F_1 and E_1 , say by a factor r , but would not affect other components we have,

$$C_{D_2}^T = F_0 + \alpha r(F_1 + E_1) + \alpha(F_2 + E_2).$$

Parameters of $C_{D_1}^T$ and $C_{D_2}^T$ were estimated as follows:

F_0, F_1, F_2, E_1, E_2 : These were estimated under D_0 from a special Time and Cost study (Lemaitre 1983), carried out as part of the redesign research program. Since the field test of D_1 revealed no discernable differences in data collection costs between D_0 and D_1 , these parameters were assumed unchanged under D_1 .

α : Field testing of telephone interviewing carried out as part of the redesign research program did not have as an objective the estimation of cost savings. An estimated 10% reduction in total data collection costs was made by Regional Operations staff, which permitted calculation of α .

r : This parameter could not be estimated based on available data, rather a Monte Carlo simulation study was needed, which is described in Appendix B.

3.4 Results of Cost-Variance Analyses

Variance Analysis: D_1 vs. D_0

Components of variance for 6 labour force characteristics were obtained for designs D_0 and D_1 using 1971 Census data for 5 ER's across Canada. Table 2 gives the % contribution from each stage of sampling to the total variance under D_0 . It can be observed that 30-40% of the total variance under D_0 was due to the rural cluster (3rd) stage of sampling, and that under design D_1 20-30% variance reductions could be obtained.

Table 2

Percent Contributions to the Total Variance from Stages of Sampling for the Current Design and Percent Reduction in the Total Variance Due to Eliminating Cluster Stage of Sampling in Rural Areas; $100 (1 - \frac{V_{D_1}}{V_{D_0}})$

Characteristic	Percent Contribution to Total Variance from						Percent Variance Reduction; $100 (1 - \frac{V_{D_1}}{V_{D_0}})$
	Urban			Rural			
	1st stage	2nd stage	3rd stage	2nd stage	3rd stage	4th stage	
LF Population	14.5	12.9	10.8	5.8	40.5	15.5	30.5
Employed	21.2	11.2	10.4	6.3	35.0	15.8	27.1
Unemployed	12.6	15.8	16.6	4.8	33.0	17.2	24.8
Not in LF	24.7	11.9	10.7	4.8	32.9	15.1	22.9
Employed Agr.	42.4	1.0	0.8	12.3	30.8	12.6	20.4
Employed Non-Agr.	23.3	12.7	11.9	5.6	31.7	14.8	21.8

The gains might be less since for the study, the variables being estimated and the size measures referred to the same point in time whereas this would not be true in practice. No attempt was made to discount the gains, however, since the choice between D_1 and D_0 was clear both in terms of variances, and on operational grounds (as discussed in Subsection 3.1). Further efforts were devoted hence to the choice between D_1 and D_2 .

Variance Analysis: D_2 vs. D_1

In this study the number of ER's was expanded to 11, and study variables (employed and unemployed) were based on the 1976 Census, whereas size measures were based on the 1971 Census. Also variances were computed with ratio estimation based on total population.

The average variance efficiency of D_2 with respect to D_1 was 1.16 for employed and 0.97 for unemployed (Table 4).

Cost Analysis: D_2 vs. D_1

Values of all the parameters in the cost model are presented in Table 3 along with $C_{D_1}^T$ and $C_{D_2}^T$ and their ratio.

As expected the between PSU and between secondary component of interviewer fees and expenses are higher under D_1 due to the frequent lack of contiguity between rural and urban portions of PSU's. The average reduction factor r in these components under D_2 was estimated as in Table 3 leading to an overall cost efficiency for D_2 vs. D_1 of 1.08 (Table 4).

Combined Cost Variance Analysis: D_2 vs. D_1

Table 4 gives the relative cost-variance efficiencies of D_2 vs. D_1 under telephone interviewing. In terms of overall efficiency, D_2 is 25% and 5% more efficient than D_1 for employed and unemployed respectively.

Based on these findings it was decided to adopt D_2 in the 2/3 of ER's capable of supporting both urban and rural strata, and design D_1 was adopted in the remaining cases.

Table 3
Values of Parameters in the NSR Cost Model and Relative Cost
Efficiencies of D_1 vs. D_2 with Telephone Interviewing

ER	F_0	F_1	F_2	E_1	E_2	α	r	$C_{D_1}^T$	$C_{D_2}^T$	$C_{D_1}^T/C_{D_2}^T$
22	2.05	0.74	1.31	0.95	0.92	0.85	0.93	5.38	5.28	1.02
32	2.13	0.86	1.11	0.90	0.97	0.84	0.88	5.35	5.17	1.03
41	2.04	0.94	0.94	0.96	0.69	0.84	0.42	5.01	4.08	1.23
44	2.04	0.94	0.94	0.96	0.69	0.84	0.50	5.01	4.21	1.19
51	1.94	0.80	1.07	0.81	0.75	0.84	0.89	4.82	4.67	1.03
56	1.94	0.80	1.07	0.81	0.75	0.84	0.68	4.82	4.39	1.10
63	2.07	1.03	1.03	1.19	0.97	0.75	0.87	5.66	5.41	1.05
72	1.92	0.96	1.13	1.05	1.09	0.85	0.82	5.52	5.21	1.06
82	1.88	1.12	1.01	1.20	0.94	0.86	0.57	5.55	4.69	1.18
86	1.88	1.12	1.01	1.20	0.94	0.86	0.90	5.55	5.35	1.04
96	2.03	0.81	1.22	0.75	0.85	0.84	0.75	5.07	4.74	1.07

Table 4
Relative Cost-Variance Efficiencies of D_1 vs. D_2

ER	Variance Efficiency V_{D_1}/D_{D_2}		Cost Efficiency $C_{D_1}^T/C_{D_2}^T$	Relative Cost-Variance Efficiency $V_{D_1}C_{D_1}^T/V_{D_2}C_{D_2}^T$	
	Employed	Unemployed		Employed	Unemployed
22	1.09	0.93	1.02	1.11	0.95
32	0.91	0.72	1.03	0.94	0.74
41	1.14	0.86	1.23	1.40	1.06
44	1.39	1.14	1.19	1.65	1.37
51	0.96	1.01	1.03	0.99	1.04
56	1.12	1.51	1.10	1.23	1.66
63	1.35	1.06	1.05	1.41	1.11
72	1.00	0.91	1.06	1.06	0.96
82	1.09	1.01	1.18	1.27	1.19
86	1.20	1.05	1.04	1.25	1.09
96	1.38	1.05	1.07	1.48	1.12
All*	1.16	0.97	1.08	1.25	1.05

* Weighted average by population size.

3.5 Special 2-Stage Design for Prince Edward Island

For Canada's smallest province, Prince Edward Island, where sampling rates of 4% are required in order to produce reliable provincial data, design alternative D_3 , a stratified sample of EA's and dwellings, was considered as an alternative to D_2 .

D_3 did not feature any clustering of the sample into geographically contiguous primaries designed to correspond to interviewers assignments, as it was hypothesized that given the high sampling rates, the increase in data collection costs might be more than offset by variance reductions due to elimination of a stage of sampling, and due to stratification gains resulting from having more strata (i.e., up to 4 times as many as under D_2).

Cost-variance study results showed the variance efficiency of D_3 vs. D_1 to be 2.39 for employed and 1.20 for unemployed, while costs under D_3 were only 8% greater. Hence, based on overall cost-variance efficiencies of 2.21 for employed and 1.11 for unemployed, D_3 was opted for.

3.6 Number of PSU's Selected Per Stratum

Under both designs D_1 and D_2 , the sample yield per PSU was fixed at 55-60 dwellings to correspond to an interviewer's assignment. In about half of the ER's, there was only enough sample for 2 or 3 PSU's to be selected. Further stratification in these cases was ruled out on the grounds that there should be at least 2 PSU's per stratum to permit unbiased estimation of variance.

For the remaining ER's, some consideration was given to having 4-5 PSU's per stratum, as this would permit greater flexibility to reduce the size of the area sample, for example, if a portion of the area sample at some time in the future were to be converted to a telephone sample under a dual frame set-up. However, stratification to the point of 2-3 PSU's per stratum was adopted, based on variance reductions of 14.8% for employed and 5.4% for unemployed for these ER's. A detailed description of the stratification procedures followed can be found in Drew, Bélanger, and Foy (1985).

4. COST-VARIANCE OPTIMIZATION BETWEEN SR and NSR AREAS

The next step in the cost-variance optimization of the LFS design was the optimization of the allocation of sample between SR and NSR areas. We used the simple cost and variance models considered by Fellegi, Gray, and Platek, (1967), i.e.,

$$\text{cost:} \quad C = \sum_{j=1}^2 C_j \frac{P_j}{W_j}, \quad (4.1)$$

$$\text{variance:} \quad V = \sum_{j=1}^2 W_j P_j \sigma_j^2, \quad (4.2)$$

where j = area type (= 1 for SR; = 2 for NSR),
 C_j = unit (i.e., per person) cost,
 P_j = population,
 $1/W_j$ = sampling rate,
 σ_j^2 = unit variance.

Fellegi et al. showed that if C is minimized with V fixed the ratio of the sampling rates is

$$\frac{W_1}{W_2} = \frac{\sigma_2}{\sigma_1} \left(\frac{C_1}{C_2} \right)^{1/2} \quad (4.3)$$

The other optimization criteria described in Section 1 also give the same ratio as above. Parameters were estimated as follows:

- (i) **Unit costs:** Historical per dwelling costs by type of area were available. These were decreased by 10% for NSR areas, to take account of the estimated effect of a shift to telephone interviewing of all rotation groups except the rotate-in group for the redesigned sample.
- (ii) **Unit variances:** Optimization was carried out with respect to the characteristic unemployed, for which variances were given by:

$$\sigma_j^2 = \beta_j \frac{u_j}{P_j} \left(1 - \frac{u_j}{P_j} \right); j = 1, 2 \tag{4.4}$$

where β_j = design effect for unemployed, and u_j = unemployed.

Historical design effects by type of area were available, and were reduced to take into account of structural improvements in the respective NSR and SR designs as described in Sections 2 and 3. Unemployment levels were based on 1980-82 average LFS data, which seemed appropriate in light of medium term forecasts which were not calling for a return to pre-1982 recession levels of unemployment, and population counts were based on the 1981 Census.

Table 5 presents the percent of sample in SR areas under the following allocations: (i) old design, (ii) proportional allocation, (iii) optimum allocation under the assumed cost and variance model, and (iv) the allocation adopted for the redesigned sample. The optimum allocation could not be adopted because of subprovincial data reliability constraints. In most cases, the differences between the optimum allocation and the one adopted are small. The optimal allocation turned out to be quite close to proportional, and quite different from the allocation under the old design.

Table 5
Percent of Sample in SR Areas within Provinces for (1) Old Sample,
(2) Proportional Allocation, (3) Optimum Allocation,
and (4) Redesign Sample

Province	Old Sample	Proportional Allocation	Optimum Allocation	Redesigned Sample
Newfoundland	41.8	51.3	42.6	44.6
Prince Edward Island	26.6	32.8	32.8	28.9
Nova Scotia	37.3	57.4	58.8	51.9
New Brunswick	49.5	52.5	47.4	53.6
Quebec	56.8	74.8	71.6	68.9
Ontario	62.5	79.1	78.8	75.0
Manitoba	54.1	71.0	76.4	56.4
Saskatchewan	44.7	51.8	62.1	56.8
Alberta	60.0	68.6	72.6	62.3
British Columbia	58.0	78.0	74.6	69.7
Canada	53.2	67.1	67.4	62.3

Table 6
 Relative Efficiency of the Redesigned Sample Allocation
 with Respect to the Old by Province (Unemployed)

Province	Cost Ratio ($= \frac{C^{(O)}}{C^{(N)}}$)	Variance Ratio ($= \frac{V^{(O)}}{V^{(N)}}$)	Rel. Eff. ($= \frac{C^{(O)}V^{(O)}}{C^{(N)}V^{(N)}}$)
Newfoundland	1.00	1.00	1.00
Prince Edward Island	1.01	1.02	1.03
Nova Scotia	1.04	1.14	1.18
New Brunswick	1.01	0.98	0.99
Quebec	1.03	1.06	1.09
Ontario	1.04	1.08	1.12
Manitoba	1.01	1.03	1.04
Saskatchewan	1.05	1.06	1.12
Alberta	1.01	1.01	1.02
British Columbia	1.02	1.09	1.11
Canada	1.03	1.07	1.10

The projected gains resulting solely from the re-allocation process under the assumption of fixed (old) provincial sample sizes and uniform sampling rates within the two area types are presented in Table 6. For this table, the unit costs and variances described above were used in determining the total costs and variances, $C^{(O)}$, $C^{(N)}$, $V^{(O)}$, $V^{(N)}$, under the old and new allocations respectively. The new allocation would have resulted in a 3% decrease in total cost and a 7% decrease in total variance of unemployed and for a combined relative efficiency (as defined in Table 6) of 1.10. Had it not been for the subprovincial data requirements, an efficiency gain of 1.12 could have been achieved under the optimal allocation.

The actual efficiency gains for the redesigned sample vs. the old sample are considered in the following section.

5. CONCLUSIONS

The changes in the LFS design taken as a result of the cost-variance studies are the following: elimination of a stage of sampling in NSR rural areas, adoption of a design featuring rural/urban stratification, adoption of a 2-stage NSR design in Prince Edward Island, increase in the number of NSR strata to the extent that only 2 or 3 PSU's per stratum will be selected, and re-optimization of the allocation of sample between NSR and SR areas. The near optimality of other design parameters established earlier by Fellegi, Gray and Platek (1967) was found to have remained unchanged, for example the number of dwellings to select per PSU in SR Areas.

The efficiency gains resulting from the changes permitted a 7% reduction in the overall LFS sample size and achieved the required reliability of subprovincial data (Singh et al. 1984) without impacting on the reliability of provincial and national estimates. The only exceptions were the provinces of Quebec and Manitoba, where greater subprovincial data demands

Table 7
Relative Efficiency of the Redesigned
vs. the Old Sample for Unemployed

Province	Cost Ratio* (= $\frac{C^{(O)}}{C^{(N)}} $)	Variance Ratio (= $\frac{V^{(O)}}{V^{(N)}} $)	Rel. Eff. (= $\frac{C^{(O)}V^{(O)}}{C^{(N)}V^{(N)}} $)
Newfoundland	1.19	1.00	1.19
Prince Edward Island	1.10	1.13	1.24
Nova Scotia	1.22	1.04	1.27
New Brunswick	1.17	0.99	1.16
Quebec	1.15	0.95	1.09
Ontario	1.13	1.03	1.16
Manitoba	1.17	0.96	1.12
Saskatchewan	1.23	1.02	1.25
Alberta**	1.15	1.00	1.15
British Columbia	1.15	1.01	1.16
Canada	1.17	0.99	1.16

* Based on the redesigned sample with telephone interviewing and the old sample with personal visit interviewing in NSR areas.
** Supplementary sample not included.

necessitated a slight loss in provincial data reliability. Table 7 gives the cost, variance and combined cost-variance ratios for the old sample (old design with 55,500 hhlds/month and no telephone interviewing in NSR's) vs. the redesigned sample (new design with 51,600 hhlds/month and telephone interviewing). The significant cost reductions are due to the shift to telephone interviewing in months 2-6 in NSR areas, and the sample size reduction. The overall cost-variance efficiency of the redesigned sample relative to the old sample was 1.16 (Table 7).

APPENDIX A

Variance Formula and Computation Method for RPPSS Sampling

Suppose that a sample of size n is selected by the randomized PPS systematic sampling from N units. Let p_i be the normalized size measure of the i -th unit such that $\sum_{i=1}^N p_i = 1$. The Horvitz-Thomson estimator of the total Y for a characteristics y is given by (Horvitz and Thomson 1952):

$$\hat{Y}_{HT} = \sum_{i \in S} \frac{y_i}{\pi_i},$$

Where S = the selected sample of size n

y_i = y -values of i -th unit

π_i = np_i , the probability that the i -th unit is in S .

and its variance is

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{i < j} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

where π_{ij} is the joint probability that both the i -th and j -th units are in S . Hartley and Rao (1962) gave an asymptotic formula for π_{ij} 's.

An exact formula by Connor (1966) is also available but quite involved. Recently Hidioglou and Gray (1980) developed a computer algorithm using a modification of Connor's formula due to Gray (1971), which was used in our study and compared with the Hartley-Rao approximation. It was found that the Hartley-Rao approximations are very close to the exact values for $N \geq 16$. We decided to use the Hidioglou-Gray algorithm for $N < 16$ and the Hartley-Rao approximation for $N \geq 16$ considering exponential increase in computation with the algorithm as N increases.

APPENDIX B

Cost Simulation of D_2 vs. D_1

In order to estimate r , the ratio of fees and expenses for travel from home to area, between PSU's, and between secondaries under NSR design alternatives D_2 and D_1 , a Monte Carlo study was carried out. The sample frames under D_1 and D_2 were simulated to the level of secondaries using Census data for each of the 11 study ER's. Fifty samples were drawn following each design, and the selected secondaries for each sample were grouped into geographically optimal assignments. If $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ are the average measures of within assignment geographic dispersion under designs D_1 and D_2 , then r was estimated by

$$\bar{M}^{(2)} / \bar{M}^{(1)}.$$

The M -measure for a given sample was defined in the following manner. Suppose that k interviewers cover an ER and $G_i = \{U_{ij}; j = 1, 2, \dots, n_i\}$ is the i -th interviewer's assignment, with n_i second stage sampling units. Let (x_{ij}, y_{ij}) be the population centroid of U_{ij} defined in Euclidean coordinates. The M -measure for the ER is defined as

$$M = \sum_{i=1}^k M_i,$$

$$M_i = \sum_{j=1}^{n_i} \{(x_{ij} - \bar{x}_i)^2 + (y_{ij} - \bar{y}_i)^2\}^{1/2},$$

where (\bar{x}_i, \bar{y}_i) is the center of G_i , i.e., $\bar{x}_i = 1/n_i \sum_{j=1}^{n_i} x_{ij}$; $\bar{y}_i = 1/n_i \sum_{j=1}^{n_i} y_{ij}$.

The determination of optimum interviewer assignments, that is the minimization of the M -measure, reduces to a classification or clustering problem. The following clustering algorithms were investigated:

i) Friedman-Rubin (1967) Transfer Algorithm

This non-hierarchical algorithm which was adopted for stratification of the LFS sample (Drew et al. 1985), starts with a random partitioning of units and proceeds towards a local optimum by moving one unit at a time from one cluster to another if the move

reduces M . It also checks that size constraints are not violated before moving a unit. An approximation to the global optimum is achieved by taking several initial random starts. A disadvantage of the Friedman-Rubin algorithm in this case was that the strict size constraints required in order to have approximately equi-sized assignments, restricted the movement of units between clusters.

ii) **Dahmström-Hagnell (1975) Exchange Algorithm**

This algorithm is similar to the Friedman-Rubin algorithm, except that it is based on exchanging pairs of units between clusters as opposed to transferring individual units. Hence it works better under strict size constraints.

iii) **Combined Algorithms**

Define a cycle of a combined algorithm as application of the exchange algorithm, followed by the transfer algorithm. Then we considered both single and two cycle combined algorithms.

The combined two cycle algorithm worked best, requiring the smallest number of random starts and the least computing cost to achieve the same level of optimality as the other algorithms. Performance of the 1 and 2 cycle combined algorithms based on 21 replicates is summarized below.

	One Cycle				Two Cycle		
	No. of Random Starts				No. of Random Starts		
	1	2	4	10	1	2	4
M-measure*	336.18	329.19	325.65	325.51	327.55	325.69	325.51
Standard Deviation	15.84	15.45	15.67	15.69	16.10	15.67	15.69
Computing Cost (\$)	5.94	11.24	21.67	53.90	8.17	15.12	29.38

* Average over 21 replicates.

REFERENCES

CONNOR, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

DAHMSSTRÖM, P., and HAGNELL, M. (1975). Multivariate stratification of primary sampling units in multi-stage sampling with an application to SCB's general purpose sample. Research Report, University of Lund.

DREW, J.D., BÉLANGER, Y., FOY, P. (1985). Multivariate clustering algorithm for stratification and its application to the Canadian Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada (in preparation).

FELLEGI, I.P., GRAY, G.B., and PLATEK, R. (1967). The new design of the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 62, 421-453.

FRIEDMAN, H.P., and RUBIN, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

- GRAY, G.B. (1971). Joint probability of selection of units in systematic samples. *Proceedings of American Statistical Association*, 271-276.
- HARTLEY, H.O., and RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- HIDIROGLOU, M.A., and GRAY, G.B. (1980). Construction of joint probability of selection for systematic PPS sampling. *Journal of Royal Statistical Society*, C29, 107-112.
- HORVITZ, D.G., and THOMSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.
- LEMAITRE, G. (1983). Some results from Time and Cost Study. Technical Report, Census and Household Survey Methods Division, Statistics Canada.
- PLATEK, R., and SINGH, M.P. (1976). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.
- RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). A simple procedure of unequal probability sampling without replacement. *Journal of Royal Statistical Society*, B24, 482-491.
- SINGH, M.P., and DREW, J.D. (1981). Research plans for the redesign of the Canadian Labour Force Survey. *Proceedings of the Section of Survey Research Methods, American Statistical Association Meetings*.
- SINGH, M.P., DREW, J.D., and CHOUDHRY, G.H. (1984). Post '81 Censal redesign of the Canadian Labour Force Survey. *Survey Methodology*, 10, 127-140.

Performance of ARIMA Models in Time Series¹

KIM CHIU, JOHN HIGGINSON, and GUY HUOT²

ABSTRACT

This study is mainly concerned with an evaluation of the forecasting performance of a set of the most often applied ARIMA models. These models were fitted to a sample of two hundred seasonal time series chosen from eleven sectors of the Canadian economy. The performance of the models was judged according to eight variable criteria, namely: average forecast error for the last three years, the chi-square statistic for the randomness of the residuals, the presence of small parameters, overdifferencing, underdifferencing, correlation between the parameters, stationarity and invertibility. Overall and conditional rankings of the models are obtained and graphs are presented.

KEY WORDS: X11-ARIMA; Ranking; Priority; Criteria

1. INTRODUCTION

Our socio-economic environment is unstable and uncertain; inflation, recessions, and increasing pollution are among the factors contributing to increasing instability. We try to resolve the problem by using a method of forecasting that permits us to evaluate the impact of the frequent changes. ARIMA models (Box – Jenkins, 1970) are flexible enough to deal with such frequent changes in time series.

The purpose of this paper is to study a set of eight criteria which when applied to the Box-Jenkins method permit an evaluation of the fitting and forecasting performance of a set of the most often applied ARIMA models to Canadian economic time series. The question of which models perform well is important for programs like the X-11-ARIMA (Dagum 1980) which automatically fits a fixed small set of models (three models in the case of the X-11-ARIMA) to the series.

Section 2 introduces eight criteria: the average forecast error for the last three years, the chi-square statistic for the randomness of the residuals, the presence of small parameters, overdifferencing, underdifferencing, correlation between the parameters, stationarity and invertibility. Section 3 discusses the criteria and summarizes the results. Section 4 ranks the models conditionally and unconditionally. Section 5 compares within-sample and out-of-sample extrapolated values for the last three years.

2. THE CRITERIA

In this section we give a brief discussion of the eight criteria used in ranking the models.

¹ Presented at (1) Business and Economic Forecasting Session of the Canadian Operational Research Symposium, Ottawa, May 1984 and (2) Business and Economic Statistics Section of the American Statistical Association Meetings, Philadelphia, August 1984.

² K. Chiu, J. Higginson, and G. Huot, Time Series Research and Analysis Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Stability

The stability condition of a process Z_t is either “stationary” or “non-stationary”. It indicates how well the system remembers the shocks a_{t-j} , $j = 1, 2, \dots$, and how fast or slowly the response of the system to any particular shock decays. For a process

$$\begin{aligned} Z_t &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots \\ &= \psi(B) a_t, \end{aligned}$$

where $a_t \sim NID(0, \sigma_a^2)$, the filter is said to be stable if the sequence $\{\psi_i\}$ is convergent. For a general ARIMA model (p, d, q) ,

$$\phi(B) (1 - B)^d Z_t = \theta(B) a_t,$$

the stability condition is that all the λ_i of the characteristic equation

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p = (1 - \lambda_1 B) (1 - \lambda_2 B) \dots (1 - \lambda_p B) = 0$$

for the process are strictly inside the unit circle, i.e. $|\lambda_j| < 1$.

Invertibility

The process Z_t may be expressed as:

$$Z_t = a_t + \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \dots$$

The system is said to be invertible if the sequence $\{\pi_i\}$ is convergent. The criterion is considered to be of primary importance because if the invertibility condition fails, the generating function $\pi(B)$ of the π 's increases without bound. This means the current event of the system depends more on events in the distant past than in the recent past, and the process is physically meaningless.

The invertibility condition for a general ARIMA model (p, d, q) , is that the ν_i of the characteristic equation

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q = (1 - \nu_1 B) (1 - \nu_2 B) \dots (1 - \nu_q B) = 0$$

for the process are strictly within the unit circle, i.e. $|\nu_i| < 1$.

Underdifferencing

In the AR(p) model, when one or more of the λ_i , say λ_k approaches 1; then from

$$\begin{aligned} \phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \\ &= (1 - \lambda_1 B) \dots (1 - \lambda_{k-1} B) (1 - \lambda_k B) \dots (1 - \lambda_p B) \\ &= (1 - \lambda_1 B) \dots (1 - \lambda_{k-1} B) (1 - \lambda_{k+1} B) \dots (1 - \lambda_p B) (1 - \lambda_k B), \end{aligned}$$

we have $\phi(B)$ approaching

$$(1 - \phi'_1 B - \phi'_2 B^2 - \dots - \phi'_{p-1} B^{p-1}) (1 - B).$$

Therefore, a differencing operator may be needed for this system, and the AR(p) model becomes an ARI($p - 1, 1$) model. Furthermore, when λ_k approaches 1, we may have non-stationarity.

Overdifferencing

Consider the general ARIMA model (p, d, q) (P, D, Q)_s,

$$\phi(B)\Phi(B)(1 - B)^d(1 - B^s)^D Z_t = \theta(B)\Theta(B)a_t.$$

If any ν_i of the characteristic equation $\theta(B) = 0$ approach 1, i.e. if any $(1 - \nu_i B)$ approach $(1 - B)$, we can eliminate $(1 - B)$ from both sides.

Test of randomness for the a_t 's

Correlation in the residuals is not desirable since we want an unbiased estimate of the parameters for the process.

The statistic

$$Q = n(n + 2) \sum_{k=1}^m (n - k)^{-1} \varrho_k^2$$

as modified by Prothero and Wallis (1976) and Ljung and Box (1978) from the Chi-square test of Box and Pierce is used.

Here n is the sample size, $k = 1, 2, \dots, m$ are the various lags, and ϱ_k are the autocorrelations. Q is used for the testing of the randomness of the residuals.

Small Parameters

Generally speaking, when the number of parameters of a given model is increased, the mean sum of squares σ_a^2 is reduced. However, only large parameters, or those parameters significantly different from 0 can contribute to a significant reduction of σ_a^2 . To check for a small parameter, we may need an F-test (Pandit and Wu 1983):

$$F = \frac{A_1 - A_0}{s} \div \frac{A_0}{N - r} \sim F(s, N - r)$$

where r is the number of parameters of the model and s is the number of parameters which are restricted to zero. N is the number of observations, A_0 is the smaller sum of squares of the restricted model, and A_1 is the larger sum of squares of the restricted model.

But in our study here, we choose two constants, 0.05 and 0.10, as our indicator of the presence of a small parameter.

Correlation of the Parameters

High positive or negative correlation between parameters reflects ambiguity in the estimated values since a range of parameter values results in models with equally good fit. Therefore, if some of the elements in the correlation matrix of estimated parameters are large in absolute value, say greater than or equal to 0.9, the model may be reduced by deleting some of the smaller parameters.

Forecasting Error

No matter how we define a good model or bad model, we still have a primary interest in the forecasting error of the model. In this paper we use the mean absolute percentage forecasting error of one-year-ahead forecast

$$MAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{Z_{t+\ell} - \hat{Z}_t(\ell)}{Z_{t+\ell}} \right| \times 100\%$$

where ℓ is 12 or 4, and $\hat{Z}_t(\ell)$ is the forecast with lead time ℓ .

3. EVALUATION OF THE ARIMA MODELS

The eight criteria have been put into two groups. The first group considers good fitting of parsimonious models while the second considers the quality of the forecasts. This distinction between fitting and forecasting is important; good fitting and good forecasting are not equivalent.

These criteria have been used to evaluate and rank seven of the most often applied ARIMA models, namely:

- | | |
|-------------------------------------|-------------------------------------|
| 1. (0, 1, 1) (0, 1, 1) _s | 5. (1, 1, 0) (0, 1, 1) _s |
| 2. (0, 1, 2) (0, 1, 1) _s | 6. (2, 1, 0) (0, 1, 1) _s |
| 3. (0, 2, 2) (0, 1, 1) _s | 7. (2, 1, 0) (0, 1, 2) _s |
| 4. (2, 1, 2) (0, 1, 1) _s | |

where “s” is 12 if the series is monthly and 4 if it is quarterly.

These models were fitted to a sample of 167 monthly seasonal time series chosen randomly from eleven sectors of the Canadian economy: national accounts; labour; prices; manufacturing; fuel, power and mining; construction; food and agriculture; domestic trade; external trade; transportation; and finance. About 40 quarterly time series from national accounts and finance were also tested.

The series are mostly multiplicative, according to the Bell Canada model test (Higginson 1976). That is, the different components (trend-cycle, seasonal, and irregular) are multiplied together to produce the raw series. Therefore, the amplitudes of the seasonal component frequently increase with increasing levels of the trend. The multiplicative series received a logarithmic transformation before the first three and last three models were fitted. The fourth model was fitted to the untransformed series in all cases.

Looking at the non-seasonal part of an ARIMA model which is associated with the trend-cycle and extremes, we see that the models can be grouped into three classes. Class I is models 1, 2 and 3 whose ordinary part includes only one or two first differences and one or two moving average parameters. Class III includes models 5, 6 and 7 whose ordinary part includes only one first difference and some autoregressive parameters. Model 4 (Class II) forms a class by itself; its non-seasonal part is mixed. We see that the seasonal part of all models is the same except for model 7.

Although the eight criteria are analysed separately in this section, several of them are dependent. For example, we shall see that the excess of parameters in model 4 generates problems of nonstationarity, noninvertibility, under- and overdifferencing, and correlation.

In Sections 3 and 4, we test within-sample extrapolated values for the seven ARIMA models. That is, the models are fitted to the whole series thus providing the parameters to be used for calculating the forecasts for the last three years. This is the way ARIMA forecasts are evaluated in the X-11-ARIMA program.

3.1 Criteria for Fitting Parsimonious ARIMA Models

The stationarity condition requires that all the roots of the autoregressive characteristic equation be inside the unit circle. We see in Table 1 that non-stationarity occurs only for model 4, in three cases. These appear to be due to overparametrization of the model.

In order for the model to be invertible, it is necessary that the roots of the moving average characteristic equation be inside the unit circle. Only model 4 has many cases of noninvertibility, 20%, as we see in Table 2. Two explanations are possible. There is first of all the case of straightforward noninvertibility. In some other cases noninvertibility was accompanied by nonstationarity. The fact that the autoregressive part may have roots near unity might have caused autocorrelation in the residuals. The moving average parameters would then take higher values to compensate.

An important criterion in judging the appropriateness of the ARIMA models for the series is the chi-square test of Box and Pierce (1970) (modified by Prothero and Wallis in 1976, and by Ljung and Box in 1978), applied to the autocorrelation of the residuals. Table 3 shows for each of the seven models the number and the percentage of series that fail the chi-square test at different levels. We see from this table first, that within a given class of models the simpler models have higher failure rates and second, that the failure rate depends to a large degree on the class of the model. The first point is illustrated by models 2 and 6 which having one more parameter than models 1 and 5, have a higher number of series passing this test. The evidence for the second point is that moving average models appear to satisfy the

Table 1
Failure in Stationarity

CRITICAL VALUE	CLASS I			CLASS II		CLASS III		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	
	(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)	(1, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 2)	
--	--	--	--	3 2%	--	--	--	--

Table 2
Failure in Invertibility

CRITICAL VALUE	CLASS I			CLASS II		CLASS III		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	
	(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)	(1, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 2)	
--	1 1%	2 1%	3 2%	33 20%	2 1%	2 1%	1 1%	

Table 3
Failure in Chi-Square

CRITICAL VALUE	CLASS I						CLASS II		CLASS III					
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1)	(0, 1, 1)	(0, 1, 2)	(0, 1, 1)	(0, 2, 2)	(0, 1, 1)	(2, 1, 2)	(0, 1, 1)	(1, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 2)
1%	31	19%	18	11%	29	17%	26	16%	62	37%	21	13%	20	12%
5%	45	27%	36	22%	46	28%	41	25%	82	49%	49	29%	42	25%
10%	61	37%	48	29%	56	34%	55	33%	89	53%	60	36%	56	34%
15%	72	43%	57	34%	69	41%	66	40%	101	60%	71	43%	64	38%
20%	83	50%	62	37%	80	48%	76	46%	106	64%	80	48%	73	44%
30%	100	60%	77	46%	94	56%	88	53%	119	71%	95	57%	89	53%
40%	111	66%	97	58%	107	64%	99	59%	127	76%	104	62%	100	60%
50%	121	72%	106	63%	118	71%	113	68%	135	81%	117	70%	116	69%
60%	131	78%	121	72%	128	77%	129	77%	141	84%	127	76%	121	72%

chi-square test better than autoregressive models. This may be due to the presence of extremes in the series. At the 5% level for example, model 1 fails for 27% of the series compared with 49% for its autoregressive counterpart model 5. As well as all models of class III, the mixed model, class II, is inferior to the second model of class I.

Underdifferencing occurs when a root of the characteristic equation of the autoregression polynomial is close to unity, say a distance ξ from unity. Here ξ is set equal to 0.1. We see in Table 4 that only model 4 is underdifferenced. This may be attributed to overparametrization. Model 4 has two autoregressive parameters and two moving average parameters in its non-seasonal part. Just through the estimation, there is a moderate chance that at least one of the autoregressive parameters will be greater than or equal to 0.9.

In this discussion the critical levels chosen for overdifferencing are 0.90 and 0.95. Table 5 shows that models 3 and 4 are most often overdifferenced. Model 3 has two first differences and two non-seasonal moving average parameters. If the second first difference is not necessary, autocorrelation is created in the series that has been differenced once already. The moving average polynomial will model this introduced autocorrelation by having one of its roots close to unity. We can therefore simplify the model by eliminating one moving average parameter and one difference. As to model 4, this may be due to overparametrization.

Table 4
Failure in Underdifferencing

CRITICAL VALUE	CLASS I						CLASS II		CLASS III					
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1)	(0, 1, 1)	(0, 1, 2)	(0, 1, 1)	(0, 2, 2)	(0, 1, 1)	(2, 1, 2)	(0, 1, 1)	(1, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 2)
.90	--	--	--	--	--	--	14	8%	--	--	--	--	--	--

In ARIMA modelling of a stochastic process, it is enough to consider the first two moments, that is, the mean and autocovariance. The test on the size of the parameters serves only to eliminate those that contribute very little or nothing to the explanation of the autocovariance.

Table 6 illustrates two things. First, the simplest models pass this test better than more complicated models. After a logarithmic transformation, most of the multiplicative series in the sample will follow a straight line fairly closely (except for seasonal variation), so a “first difference” model will fit them using few parameters. Adding an extra unnecessary parameter to the model will often result in its receiving a small estimate from the estimation. Second, the estimated values of the moving average parameters are small (less than .05 or .10) more often than the estimated values of the autoregressive parameters. For example at the level of 0.05, the second autoregressive parameter in model 6 is judged unnecessary 13% of the time compared with 29% of the time for the second moving average parameter in model 2. Similarly, the addition of a second seasonal moving average parameter increased the failure rate from 13% in model 6 to 43% in model 7.

Table 5
Failure in Overdifferencing

CRITICAL VALUE	CLASS I						CLASS II		CLASS III			
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	(0, 1, 1)	(0, 1, 1)	(0, 1, 2)	(0, 1, 1)	(0, 2, 2)	(0, 1, 1)	(2, 1, 2)	(0, 1, 1)	(1, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 1)
.90	8	5%	11	7%	43	26%	50	30%	7	4%	9	5%
.95	3	2%	6	4%	19	11%	37	22%	3	2%	3	2%

Table 6
Failure in Small Parameter

CRITICAL VALUE	CLASS I						CLASS II		CLASS III			
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	(0, 1, 1)	(0, 1, 1)	(0, 1, 2)	(0, 1, 1)	(0, 2, 2)	(0, 1, 1)	(2, 1, 2)	(0, 1, 1)	(1, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 1)
.05	15	9%	49	29%	21	13%	42	25%	12	7%	22	13%
.10	26	16%	88	53%	43	26%	73	44%	31	19%	45	28%

Table 7
Failure in Correlation

CRITICAL VALUE	CLASS I						CLASS II		CLASS III			
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	(0, 1, 1)	(0, 1, 1)	(0, 1, 2)	(0, 1, 1)	(0, 2, 2)	(0, 1, 1)	(2, 1, 2)	(0, 1, 1)	(1, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 1)
--	--	--	3	2%	86	51%	124	74%	--	--	--	--

High positive or negative correlations between parameter estimates are undesirable and reflect ambiguity in the estimation situation since a range of parameter combinations result in models with equally good fits. Table 7 shows that only models 2, 3 and 4 fail the correlation test, i.e. the absolute value of at least one of the correlations is ≥ 0.90 . The problem is minimal for model 2, and serious for models 3 and 4 where 51% and 74% of the fits had highly correlated parameters. This may be due to overdifferencing in model 3 and the presence of too many parameters in model 4.

3.2 Criterion for Extrapolation of ARIMA Models

This criterion attempts to ensure the quality of the forecasts of the ARIMA models. We require that the average percentage forecast error of the fitted error be below a certain level.

Table 8 shows that six of the seven models are equivalent from the point of view of forecasts, i.e. the number of autoregressive and moving average parameters does not affect the forecast error of the model averaged over all the series. Of course, some models perform better for certain series.

Table 9 shows the average forecast error and standard deviation of the error under two possible outcomes: passing and failing the forecast error criterion. Not only is the failure rate of model 3 higher than that of the other models, but the table shows that when it fails,

Table 8
Failure in Forecast Error

CRITICAL VALUE	CLASS I						CLASS II				CLASS III					
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7			
	(0, 1, 1)	(0, 1, 1)	(0, 1, 2)	(0, 1, 1)	(0, 2, 2)	(0, 1, 1)	(2, 1, 2)	(0, 1, 1)	(1, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 2)		
%	%		%		%		%		%		%		%			
10	89	53	84	50	101	60	80	48	84	50	85	51	85	51		
15	57	34	58	35	69	41	53	32	57	34	56	34	55	33		
20	39	23	40	24	51	31	40	24	40	24	40	24	40	24		
25	32	19	33	20	43	26	32	19	36	22	14	20	34	20		
30	24	14	26	16	35	21	24	14	27	16	27	16	27	16		

Table 9
Conditional Mean (M) and Standard Deviation (SD)
of the Average Forecast Error

Critical Value	Out-come	CLASS I						CLASS II		CLASS III					
		Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
		(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
15%	Pass	7%	4.0	6%	3.9	7%	4.1	6%	3.8	7%	3.9	7%	4.0	7%	3.9
	Fail	35%	22.3	36%	22.5	41%	26.4	36%	21.4	38%	24.5	37%	23.4	37%	23.0

its average forecast error is bigger. The forecast errors of model 3 are increased by its over-differencing. However, when the forecast errors of model 3 pass the criterion, their average is as small as that of the other models.

4. RANKING OF THE MODELS

To rank the models, the eight criteria are used at different acceptance levels. Tables 10 and 11 present the overall and conditional rankings of the models. Table 10 gives the total

Table 10
Overall Ranking of the Models

2 criteria FE ≤ 15% χ ² ≥ 5%		8 criteria* FE ≤ 15% χ ² ≥ 5% SP ≤ .10 OD ≥ .90		8 criteria* FE ≤ 15% χ ² ≥ 5% SP ≤ .05 OD ≥ .90		8 criteria* FE ≤ 15% χ ² ≥ 5% SP ≤ .05 OD ≥ .95	
Models	% of series that passed	Models	% of series that passed	Models	% of series that passed	Models	% of series that passed
4	52%	1	34%	6	38%	6	39%
7	51%	6	31%	1	37%	1	38%
6	49%	5	23%	2	29%	2	29%
2	48%	2	20%	5	26%	5	28%
1	44%	3	13%	7	25%	7	27%
3	41%	7	11%	3	17%	3	19%
5	32%	4	2%	4	4%	4	5%

*As well as the four criteria listed, the four other criteria mentioned in the text were imposed.

Table 11
Conditional Ranking of the Models

2 criteria FE ≤ 15% χ ² ≥ 5%		8 criteria* FE ≤ 15% χ ² ≥ 5% SP ≤ .10 OD ≥ .90		8 criteria* FE ≤ 15% χ ² ≥ 5% SP ≤ .05 OD ≥ .90		8 criteria* FE ≤ 15% χ ² ≥ 5% SP ≤ .05 OD ≥ .95	
Models	% of series that passed	Models	% of series that passed	Models	% of series that passed	Models	% of series that passed
4	52%	1	34%	6	38%	6	39%
7	9%	3	6%	3	9%	3	9%
2	1%	6	4%	7	4%	1	4%
3	1%	5	2%	2	3%	4	2%

*As well as the four criteria listed, the four other criteria mentioned in the text were imposed.

success rate of the models. Table 11 gives first the total success rate of the best model; the following models are chosen according to their success with series with which all higher models have failed.

Table 10 shows that:

- when only the chi-square statistic (χ^2) and average forecast error (FE) are used as criteria, models 4 and 7, which have the most parameters, rank at the top.
- on the other hand, the use of all criteria favour the simplest models (models 1 and 6), at all levels of small parameter (SP) and overdifferencing (OD) criteria.
- models 1 and 6 usually rank close together, although model 1 has one less parameter than model 6.
- when model 6 is not first it is a close second.
- the more the criteria are relaxed, the higher the pass ratio is, although the ranking of the models remains about the same.

In table 11 we see that:

- when all criteria are used, models 1 and 6 which ranked first and second in table 10 now rank only first and third.
- second place belongs to model 3. This model, which in table 10 ranked third, fifth and sixth with total success rates of 41%, 13%, 17%, and 19%, here ranks fourth once and second three times. This is because model 3 fits well an important family of series (series with a steep trend) that all other models fit poorly.
- moving average and autoregressive models are not mutually exclusive. These two families of models are complementary and necessary in fitting and forecasting series.
- when we require only that the average forecast error be less than 15% and the chi-square statistic be greater than 5% and nothing else, the combined success rate of models 4, 7, 2 and 3 together is 63%.
- when all the criteria are used, the models chosen are simple and their combined success rate varies between 46% and 54% using the levels of 15% and 5% described just above. The success rate depends on the levels of small parameter and overdifferencing used.

Even though model 1 does not appear in the third column of table 11, it would appear there if the level of forecast error permitted were raised to 20%.

The criteria and levels used in selecting models in figures 1 and 2 are the same as are used in the second column of tables 10 and 11, except that in figure 1 the average forecast error permitted varies between 10% and 99% while in figure 2 the chi-square criteria varies between 10% and 60%.

Figure 1 shows that:

- models 1, 3 and 6 perform the best.
- the ranking of the models tends to remain the same.
- the performance of the first model increases more rapidly than that of the others, going from 23% to 59% compared with an increase from 13% to 17% for model 3. This point needs clarification. Model 1 is chosen according to its unconditional performance, while the other models are chosen according to their conditional ranking.
- the increase in performance of the models according to unconditional ranking is greater than the increase when using conditional ranking.

We see in figure 2 that

- models 1, 3 and 6 are generally the best models for any level of chi-square.
- models 1 and 6 trade places but are not mutually exclusive.

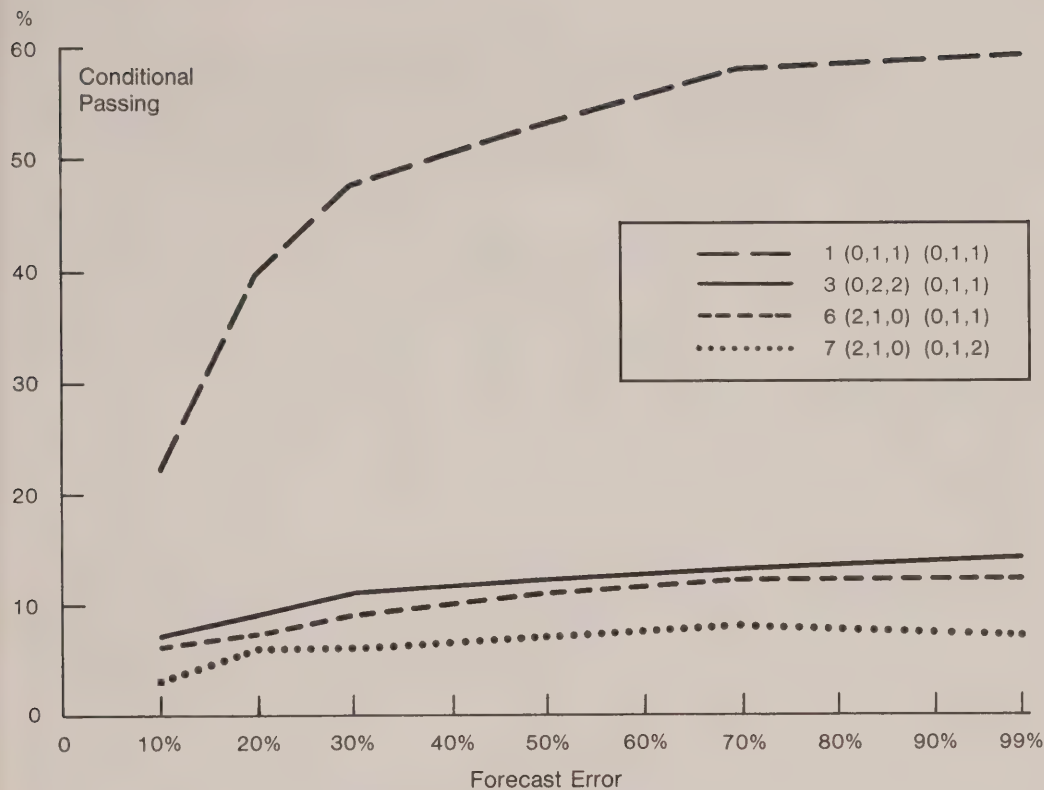


Figure 1. Model Priority Chart for Different Levels of the Forecast Criterion

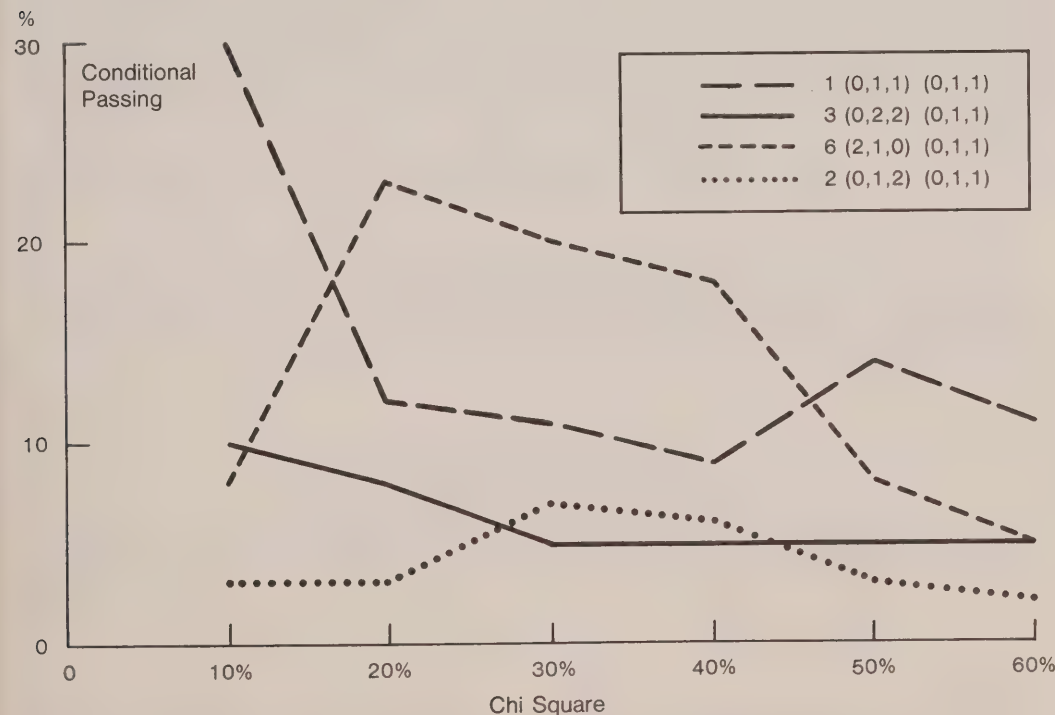


Figure 2. Model Priority Chart for Different Levels of the Chi-Square Criterion

Table 12
Conditional ranking of the ARIMA models for the sectors of the Canadian economy

Sectors	Models ranking and % of series that passed							
	first model	%	second model	%	third model	%	fourth model	%
Labour	1	79	3	14	–	0	–	0
Prices	5	50	7	17	2	8	–	0
Manufacturing.....	3	19	6	14	1	5	2	5
Fuel, Power and Mining	1	46	6	4	–	0	–	0
Domestic Trade.....	1	53	6	7	7	7	–	0
External Trade	6	21	–	0	–	0	–	0
Transportation	1	54	5	8	–	0	–	0
Finance	1	32	3	11	–	0	–	0

Table 12 presents the conditional ranking of the ARIMA models for those sectors of the Canadian economy for which we fitted twelve or more series. The criteria and levels used in ranking the models are the same as those used in the second column of tables 10 and 11. We see that

- models 1 and 6 are generally the best performers.
- the combined success rate of the models varies considerably from one sector to another, from 93% in the labour sector to only 21% in external trade.
- this success rate is at least 50% for five sectors. The rate depends on the structure of the series, changes in the structure, and the amount of irregular in the series. The rate is good considering that for two of the last three years Canada suffered a severe recession which strongly affected the structure of the series. The success rate for external trade is always low because those series are very irregular.

5. WITHIN-SAMPLE AND OUT-OF-SAMPLE FORECASTS

The within-sample forecasts are obtained by fitting the models to the entire series in order to estimate the parameters and calculate the forecasts for the last three years. The out-of-sample forecasts do not use information from after the forecast time origin. For each forecast origin, the parameters are re-estimated.

Table 13
Failure Rate in Forecast Error for
Within-Sample and Out-of-Sample Forecasts

	Model 1 (0, 1, 1) (0, 1, 1)	Model 2 (0, 1, 2) (0, 1, 1)	Model 3 (0, 2, 2) (0, 1, 1)	Model 4 (2, 1, 2) (0, 1, 1)	Model 5 (1, 1, 0) (0, 1, 1)	Model 6 (2, 1, 0) (0, 1, 1)	Model 7 (2, 1, 0) (0, 1, 2)
	%	%	%	%	%	%	%
Within-sample	34	35	41	32	34	34	33
Out-of-sample	31	32	42	33	31	32	31

Table 14
Conditional and Unconditional Ranking of the Models

Unconditional ranking		Conditional Ranking	
Models	% of series that passed	Models	% of series that passed
1	40%	1	40%
6	28%	2	5%
5	27%	7	4%
2	20%	3	3%
3	14%		
7	10%		
4	2%		

Table 13 shows the rate of failure in forecast error at the 15% level for within-sample and out-of-sample forecasts. The difference between the two is small and is well within one standard deviation for each model. The X-11-ARIMA seasonal adjustment program uses within-sample forecasts because they cost less.

Table 14 has been prepared using the same criteria and levels as were used in the second columns of tables 10 and 11. The unconditional ranking is exactly the same as that in the second column of table 10. Only the success rates of the first three models differ, and in table 14, model 1 is clearly superior to the other models. However, the conditional ranking is different from that appearing in the second column of table 11.

The conditional rankings in tables 11 and 14 differ for two reasons. First, of course, table 14 uses out-of-sample forecasts. Another important reason is that the calculation of the seven other criteria was based on one year less data, and the missing year contained a severe recession. Thus the structure of the series and the choice of models is markedly different.

It appears therefore that the conditional ranking of the models for both within-sample and out-of-sample forecasts depends on the phase of the business or economic cycle in which the series ends.

6. CONCLUSION

Our objective was to rank a set of seven ARIMA models according to their fitting and forecasting of a large sample of time series.

- when only the chi-square statistic and the average forecast error are used as criteria, models 4 and 7 rank at the top.
- The use of all eight criteria favours the simplest models (1 and 6) and model 3.
- Models 1 (moving average model) and 6 (autoregressive model) rank close together in unconditional ranking, although model 1 has one less parameter than model 6.
- In conditional ranking, these two both rank highly but are not mutually exclusive. That is, moving average and autoregressive models are complementary and both are necessary in fitting and forecasting series.
- Although Model 3 ranks near the bottom, it fits well an important family of series (series with a steep trend) that all other models fit poorly.
- The nonparsimonious models (numbers 4 and 7) have a combined success rate of 61% compared to a success rate that varies between 44% and 52% for parsimonious models 1, 6 and 3.

- The combined success rate of the models varies considerably from one economic sector to another, from 93% in the labour sector to only 21% in external trade. This rate depends on the structure of the series, changes in the structure, and the amount of irregular in the series.
- It appears that the conditional ranking of the models for both within-sample and out-of-sample forecasts depends on the phase of the business or economic cycle in which the series ends.

ACKNOWLEDGEMENT

We acknowledge helpful discussions with Mr. Normand Laniel and we are very grateful to Ms Helen Lim and Mr. Alfred Papineau for their valuable computational assistance, and to Ms B. Cohen for her typing assistance.

REFERENCES

- BOX, G.E.P., and JENKINS, G.M. (1970). *Times Series Analysis Forecasting and Control*. Holden Day: San Francisco.
- BOX, G.E.P., and PIERCE, D.A. (1970). Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *Journal of the American Statistical Association*, 65, 1509-1526.
- DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method*. Catalogue No. 12-564E, Statistics Canada, Ottawa.
- DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis*. John Wiley & Sons, Inc.
- HIGGINSON, J. (1976). A test for the presence of seasonality and a model test. Research Paper, Time Series Research and Analysis Division. Statistics Canada, Ottawa.
- LJUNG, G.M., and BOX, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-307.
- PANDIT, S.M., and WU, S.M. (1983). *Time Series and System Analysis with Applications*. John Wiley & Sons, Inc.
- PLOSSER, C.I., and SCHWERT, G.W. (1977). Estimation of a non-invertible moving average process. *Journal of Econometrics*, 6, 199-224.
- PROTHERO, D.L., and WALLIS, K.F. (1976). Modelling macroeconomic time series (with discussion). *Journal of the Royal Statistical Society*, A39, 468-500.

An Empirical Study of Some Regression Estimators for Small Domains

M.A. HIDIROGLOU and C.E. SÄRNDAL¹

ABSTRACT

The synthetic estimator (SYN) has been traditionally used to estimate characteristics of small domains. Although it has the advantage of a small variance, it can be seriously biased in some small domains which depart in structure from the overall domains. Särndal (1981) introduced the regression estimator (REG) in the context of domain estimation. This estimator is nearly unbiased, however, it has two drawbacks; (i) its variance can be considerable in some small domains and (ii) it can take on negative values in situations that do not allow such values.

In this paper, we report on a compromise estimator which strikes a balance between the two estimators SYN and REG. This estimator, called the modified regression estimator (MRE), has the advantage of a considerably reduced variance compared to the REG estimator and has a smaller Mean Squared Error than the SYN estimator in domains where the latter is badly biased. The MRE estimator eliminates the drawback with negative values mentioned above. These results are supported by a Monte Carlo study involving 500 samples.

KEY WORDS: Small domains; regression estimation; modified regression estimator; bias; mean squared error.

1. INTRODUCTION

The synthetic estimator (SYN) has the advantage of a small variance, but the following disadvantages: (a) it can be badly biased in some domains, and ordinarily we do not know which ones; (b) consequently, a calculated coefficient of variation (cv), or a calculated confidence interval, is meaningless for such domains.

For the same model that underlies the SYN estimator one can create a nearly unbiased analogue, the generalized regression estimator (REG), which has the additional advantage that a standard design based confidence interval is easily computed for each domain estimate. A disadvantage with REG is that the estimated variance (and hence the cv and the width of the confidence interval) can be unacceptably large in very small domains. (This is, of course, a direct consequence of the shortage of observations in such domains.) Also, the REG can (although with small probability) take negative values in situations where such values are unacceptable.

It is therefore desirable to strike a balance between SYN and REG. Here, we report on an empirical study with one such compromise estimator, the modified regression estimator (MRE). It has a small (but noticeable) bias in those domains where the synthetic estimator is greatly biased; in other domains, the MRE is nearly unbiased. The MRE has the advantage of a considerably reduced variance compared to the REG estimator. In addition, the MRE has a smaller Mean Squared Error than the SYN estimator in domains where the latter is badly biased. Meaningful confidence intervals can also be easily constructed for the new MRE estimator.

¹ M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada, 5-C8, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6 and C.E. Särndal, Department of Mathematics and Statistics, University of Montréal, Montréal, Québec, Canada H3C 3J7.

The paper is structured as follows. In Section 2, some of the commonly used estimators for small areas such as the direct, post-stratified and synthetic estimators are reviewed as well as some of the regression estimators given by Särndal (1981, 1984). In Section 3, the proposed modified regression estimators are introduced and discussed. In Section 4, the properties of the modified regression estimators as well as some of the other estimators are studied through a Monte Carlo simulation using business tax data. Finally, Section 5 provides some general conclusions.

2. ESTIMATORS

Let the population $U = \{1, \dots, k, \dots, N\}$ be divided into D non-overlapping domains $U_1, \dots, U_d, \dots, U_D$. Let N_d be the size of U_d . (In our empirical study, the domains are defined by a cross-classification of 4 industrial groupings with the 18 census divisions in the province of Nova Scotia. There were $D = 70$ non-empty domains, as described in Hidirolou, Morry, Dagum, Rao and Särndal (1984).)

The population is further divided along a second dimension, into G non-overlapping groups, $U_{.1}, \dots, U_{.g}, \dots, U_{.G}$.

The size of $U_{.g}$ is denoted $N_{.g}$. (In our study, the groups are based on Gross Business Income classes.) The cross-classification of domains and groups gives rise to DG population cells U_{dg} ; $d = 1, \dots, D$; $g = 1, \dots, G$. Let N_{dg} be the size of U_{dg} .

Then the population size N can be expressed as

$$N = \sum_{d=1}^D N_d = \sum_{g=1}^G N_{.g} = \sum_{d=1}^D \sum_{g=1}^G N_{dg} \quad (2.1)$$

Let s denote a sample of size n drawn from U by simple random sampling (srs). Denote by s_d , s_g and s_{dg} the parts of s that happen to fall, respectively, in U_d , $U_{.g}$ and U_{dg} .

The corresponding sizes, which are random variables, are denoted by n_d , n_g and n_{dg} . Note that (2.1) holds for lower case n 's as well. The variable of interest, y (= Wages and Salaries) takes the value of y_k for the k :th unit (= unincorporated business tax filer). The auxiliary variable x (= Gross Business Income) takes the value x_k for the k :th unit, and x_k is known for all $k = 1, \dots, N$.

The following estimators of the domain total $t_d = \sum_{U_d} y_k$ are compared, where \sum_{U_d} denotes the summation over the units in U_d .

The straight expansion estimator (EXP):

$$\hat{t}_{d\text{EXP}} = \frac{N}{n} \sum_{s_d} y_k \quad (2.2)$$

The poststratified estimator (POS):

$$\hat{t}_{d\text{POS}} = N_d \bar{y}_{s_d} \quad (2.3)$$

where

$$\bar{y}_{s_d} = \sum_{s_d} \frac{y_k}{n_d}$$

is the mean of the n_d y -values from the d :th domain. If $n_d = 0$ we define the POS estimator to be zero (somewhat arbitrarily, since strictly speaking the estimator is then undefined). Neither the EXP nor the POS estimator are particularly advantageous. They serve mainly as benchmarks against which the behaviour of the following more efficient estimators will be compared.

Two versions of the SYN and REG have been investigated, the "Count" version and the "Ratio" version. The SYN estimator is based on the assumption that a given model holds for each group g . For the "Count" version a given model would lead to the assumption that the mean of each group is the same across all domains d . For the "Ratio" version, the implied model would be that the ratios of a given variable of interest over an auxiliary variable would be constant within a given group across all domains. If the assumption of homogeneity of domain characteristics does not hold within each group, the SYN estimators can be very biased. The REG estimation method as given by Särndal (1984) is motivated by the following requirements: (a) to obtain approximately design-unbiased estimates with simple variance estimates and easily calculable (and meaningful) confidence intervals; (b) to strengthen the estimates by involving sample data from all domains.

The formulas for the "Count" versions are:

Synthetic-Count estimator (SYN/C):

$$\hat{t}_{dSYN/C} = \sum_{g=1}^G N_{dg} \bar{y}_{s,g} \tag{2.4}$$

where $\bar{y}_{s,g}$ is the mean of y in s_g .

Regression-Count estimator (REG/C):

$$\hat{t}_{dREG/C} = \sum_{g=1}^G \{N_{dg} \bar{y}_{s,g} + \hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s,g})\} \tag{2.5}$$

where $\bar{y}_{s_{dg}}$ is the mean of y in s_{dg} , and $\hat{N}_{dg} = Nn_{dg}/n$. Here, $\sum_{g=1}^G \hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s,g})$ is a bias correction term that ordinarily carries a considerable variance contribution.

The "Ratio" versions of the SYN and REG estimators are:

Synthetic-Ratio estimator (SYN/R):

$$\hat{t}_{dSYN/R} = \sum_{g=1}^G X_{dg} \hat{R}_g \tag{2.6}$$

with $X_{dg} = \sum_{U_{dg}} x_k$ and

$$\hat{R}_g = \frac{\sum_{s,g} y_k}{\sum_{s,g} x_k}$$

Regression - Ratio estimator (REG/R):

$$\hat{t}_{dREG/R} = \sum_{g=1}^G \{X_{dg} \hat{R}_g + \hat{N}_{dg}(\bar{y}_{s_{dg}} - \hat{R}_g \bar{x}_{s_{dg}})\} \tag{2.7}$$

3. MODIFIED REGRESSION ESTIMATORS

Regression estimators introduced by Särndal (1984) were constructed by fitting a regression model to some auxiliary variables and using the resulting fitted model to create predicted values for the units in the population domain. Assuming that the sampling design, p , is an arbitrary one (not necessarily srs) with inclusion probabilities π_k (first order) and π_{kt} (second order), let the regression model be given by

$$E_{\xi}(y_k) = x_k' \beta; \quad V_{\xi}(y_k) = v_k$$

where the y_k are independent random variables. An estimator of β is

$$\hat{\beta} = \left(\sum_s \frac{x'_k x_k}{\nu_k \pi_k} \right)^{-1} \sum_s \frac{x'_k y_k}{\nu_k \pi_k}$$

where it is assumed that the ν_k are known to multiplicative constant(s) that cancel when $\hat{\beta}$ is derived.

Following Särndal (1984), a nearly unbiased estimator of the unknown d -th domain total is given by

$$\hat{t}_{d\text{REG}} = \sum_{U_d} \hat{y}_k + \sum_{s_d} \frac{e_k}{\pi_k} \quad (3.1)$$

where $\hat{y}_k = x'_k \hat{\beta}$ is the k -th predicted value and $e_k = y_k - \hat{y}_k$ denotes the k -th residual.

We shall refer to $\sum_{U_d} \hat{y}_k$ as *the synthetic term* of the estimator $\hat{t}_{d\text{REG}}$ and the second term, $\sum_{s_d} e_k/\pi_k$, will be called the *correction term*.

If s_d is non-empty, an approximately unbiased alternative to the REG estimator (3.1) is given by

$$\hat{t}_{d\text{ALT}} = \sum_{U_d} \hat{y}_k + N_d \frac{\sum_{s_d} \frac{e_k}{\pi_k}}{\hat{N}_d} \quad (3.2)$$

where

$$\hat{N}_d = \sum_{s_d} \frac{1}{\pi_k}$$

is the estimated domain size.

The correction term now appears in the form of a ratio estimator,

$$\frac{\sum_{s_d} \frac{e_k}{\pi_k}}{\sum_{s_d} \frac{1}{\pi_k}},$$

multiplied by the known domain size N_d . (obviously, N_d is known since the cell counts N_{dg} are known).

The size n_d being random, the ratio form will serve to reduce the variance of the correction term. The effect will be particularly noticeable in domains where the average of the residuals is clearly away from zero (that is, in domains where the model does not fit well).

If the expected sample take in the domain, $E_d = E_p(n_d) = \sum_{U_d} \pi_k$, were substantial (say, $E_d \geq 50$), then it is practically certain that the realized sample take, n_d , will not be exceedingly small. For example, under srs, values $n_d \leq 30$ will hardly ever occur. In such situations, the nearly unbiased estimator (3.2) can be recommended as is. It should realize important efficiency gains over (3.1), notably in domains where the model does not fit as well. But in practice one often encounters domains that are so small that the expected sample take E_d does not exceed 5. This is true for a number of domains in our study. In such cases, realized sample takes n_d between zero and five are very likely. Our empirical work has confirmed the intuitively obvious fact that the residual correction will, in these small domains, contribute greatly to the variance, whether the correction appears in its straight form, $\sum_{s_d} e_k/\pi_k$, as in (3.1), or in its ratio form, $N_d(\sum_{s_d} e_k/\pi_k)/(\sum_{s_d} 1/\pi_k)$, as in (3.2).

To counteract this inflated variance contribution, we modify the correction term of (3.2) in a way implying that we settle for a small bias (in domains where the model fits less well) in exchange for a reduced variance contribution when the realized sample take n_d is lower than expected (and it is assumed that the expected sample take is already low in itself).

The form of the new correction term will be determined by the relation between realized sample take n_d , and expected sample take E_d . The correction term $\sum_{s_d} e_k/\pi_k$ will be multiplied by (\hat{N}_d/N_d) when $n_d < E_d$ and by (N_d/\hat{N}_d) otherwise. The resulting correction term using this adaptive “dampening factor” will have the effect of not “over-correcting” the synthetic term when some of the residuals e_k behave as outliers for small n_d ’s. The “over-correcting” may have the effect of greatly underestimating a domain d , yielding negative values when only positive values are acceptable, or conversely greatly overestimating the domain.

The resulting estimator, the modified regression estimator (MRE), incorporating these two types of realizations of n_d , is

$$\hat{t}_{dMRE} = \sum_{U_d} \hat{y}_k + F_d \sum_{s_d} \frac{e_k}{\pi_k} \tag{3.3}$$

where

$$F_d = \begin{cases} \frac{N_d}{\hat{N}_d} & \text{when } n_d \geq E_d \\ \frac{\hat{N}_d}{N_d} & \text{when } n_d < E_d \end{cases}$$

It can be shown that (3.3) is nearly unbiased conditionally on n_d , as long as $n_d \geq E_d$. For $n_d < E_d$, the MRE has some conditional bias, which tends to increase the more n_d falls short of its expected value. At the same time, the MRE estimator is being pushed towards its synthetic term, thus benefitting from the stability (low variance) of the synthetic term. Unconditionally, the MRE estimator given by (3.3) will have a certain small bias, but a much reduced variance compared with the REG estimator.

We note a final point in favour of MRE estimator. As a result of its considerable variance in very small domains, the REG estimator will, with a small but positive probability, take values extremely removed from the true value t_d . The value of the REG may even be negative, which is, of course, unacceptable for a variable (such as Wages and Salaries) which is by definition non-negative. Negative values of the REG estimate can occur when there exists large negative residuals e_k in the correction term of (3.1), and are especially likely when $n_d < E_d$. The new MRE estimator virtually eliminates this occurrence of negative estimates. In practice, if by a remote possibility the MRE takes a negative value, we recommend to redefine the MRE estimator as being equal to the always positive SYN estimator.

A natural formula for estimating the variance of (3.2) is

$$\hat{V}_p(\hat{t}_{dALT}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{\substack{k \neq \ell \\ \in s_d}} \Delta_{k\ell} \frac{(e_k - \bar{e}_{s_d})(e_\ell - \bar{e}_{s_d})}{\pi_k \pi_\ell} \tag{3.4}$$

where

$$\bar{e}_{s_d} = \frac{\sum_{s_d} \frac{e_k}{\pi_k}}{\sum_{s_d} \frac{1}{\pi_k}}$$

and

$$\Delta_{k\ell} = \begin{cases} 1 - \pi_k & \text{if } \ell = k \\ 1 - \frac{\pi_k \pi_\ell}{\pi_{k\ell}} & \text{if } \ell \neq k. \end{cases}$$

We propose that the same formula may serve well to estimate the variance of the MRE estimator (3.3). It is true that (3.3) differs from (3.2) when the realized sample take falls short of the expected sample take; however, it is not foreseen that the difference will be great enough to cause serious distortion in the validity of a confidence interval for t_d centred on $\hat{t}_{d\text{MRE}}$ using (3.4) as the estimated variance.

In the case of simple random sampling, and assuming for $g = 1, \dots, G$,

$$E_{\xi}(y_k) = \beta_g; V_{\xi}(y_k) = \sigma_g^2; k \in U_{.g}, \quad (3.5)$$

we find

$$\hat{\beta}_g = \frac{\sum_{s.g} y_k}{n_{.g}} = \bar{y}_{s.g},$$

leading to the "Count estimator" whose modified version (MRE/C) is

$$\hat{t}_{d\text{MRE/C}} = \sum_{g=1}^G \{N_{dg} \bar{y}_{s.g.} + F_d \hat{N}_{dg} (\bar{y}_{s_{dg}} - \bar{y}_{s.g.})\} \quad (3.6)$$

where E_d in the formula for F_d is now given by

$$E_d = E_{\text{srs}}(n_d) = \frac{nN_d}{N}$$

with

$$\hat{N}_{dg} = n_{dg} \left(\frac{N}{n} \right)$$

and

$$\bar{y}_{s_{dg}} = \begin{cases} \frac{\sum_{s_{dg}} y_k}{n_{dg}} & \text{for } n_{dg} \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The MRE/C estimator will have some bias, which is, however, ordinarily much less than that of the SYN/C estimator.

The underlying model assumptions which lead to the "ratio estimator", whose modified version is denoted as MRE/R, are for $g = 1, \dots, G$,

$$E_{\xi}(y_k) = \beta_g x_k; V_{\xi}(y_k) = \sigma_g^2 x_k, k \in U_{.g}.$$

The MRE/R estimator is then, in the case of simple random sampling,

$$\hat{t}_{d\text{MRE/R}} = \sum_{g=1}^G \{X_{dg} \hat{R}_g + F_d \hat{N}_{dg} (\bar{y}_{s_{dg}} - \hat{R}_g \bar{x}_{s_{dg}})\} \quad (3.7)$$

where

$$\hat{R}_g = \frac{\sum_{d=1}^D \hat{N}_{dg} \hat{y}_{s_{dg}}}{\sum_{d=1}^D \hat{N}_{dg} \hat{x}_{s_{dg}}},$$

and

$$X_{dg} = \sum_{U_{dg}} x_k.$$

Drew, Singh and Choudhry (1982) provided small domain estimators which, although not derived by a regression approach, have some similarity to the ones given in this paper. Their “count” version is

$$\hat{t}_{d\text{KNO/C}} = \sum_g N_{dg} \{ W'_{dg} \hat{y}_{s_{dg}} + (1 - W'_{dg}) \hat{y}_{s_g} \} \tag{3.8}$$

while their “ratio” version is

$$\hat{t}_{d\text{KNO/R}} = \sum_g X_{dg} \left\{ W'_{dg} \frac{\hat{y}_{s_{dg}}}{\hat{x}_{s_{dg}}} + (1 - W'_{dg}) \frac{\hat{y}_{s_g}}{\hat{x}_{s_g}} \right\} \tag{3.9}$$

where

$$W'_{dg} = \begin{cases} \frac{n_{dg}}{E_{dg}} & \text{if } n_{dg} \leq E_{dg} \\ 1 & \text{otherwise} \end{cases}$$

with $E_{dg} = n(N_{dg}/N)$. In the present context, if W'_{dg} in (3.8) is replaced by

$$W''_{dg} = \begin{cases} \left(\frac{n_d}{E_d} \right) \left(\frac{n_{dg}}{E_{dg}} \right) & \text{if } n_d < E_d \\ \left(\frac{E_d}{n_d} \right) \left(\frac{n_{dg}}{E_{dg}} \right) & \text{if } n_d \geq E_d \end{cases}$$

we obtain $\hat{t}_{d\text{MRE/C}}$.

4. RESULTS FROM THE EMPIRICAL STUDY

In order to study the properties of the estimators discussed in the preceding sections, a simulation was undertaken. The province of Nova Scotia was chosen as our population with $N = 1678$ sampling units (unincorporated tax filers). The variable of interest, y , is Wages and Salaries. We use a single auxiliary variable, x , namely, Gross Business Income. It is assumed that x_1, \dots, x_N are known.

Domains of the population were formed by a cross-classification of four industrial groups with eighteen regions. The industrial groups were Retail (515 units), Construction (496 units), Accommodation (114 units) and Others (553 units). The overall correlation coefficients between Wages and Salaries and Gross Business Income were 0.42 for Retail, 0.64 for Construction, 0.78 for Accommodation and 0.61 for Others. The regions were the 18 Census Divisions of the province. This produced 70 non-empty domains (out of the four times 18 domains, two combinations had no units). Thus, 70 domain totals t_d are to be estimated every time a sample is drawn.

For the Monte Carlo simulation, 500 simple random samples, s , each of size $n = 419$, were selected from the population of $N = 1678$ units. The selected sample units were classified into type of industry and Census Division. The population could have been divided along a second dimension, say income groups. But for the purposes of this study, all the taxfilers were considered as belonging to one income group ($G = 1$).

The results are summarized for each small area within the industrial groups RETAIL and ACCOMMODATION using tables and graphs. For the tables (1-4), summary statistics are the relative conditional bias and mean squared error. The eight graphs, one for each of the eight estimators, are given in figure 1. In each graph, there are eighteen vertical 'distribution bands', one for each of the eighteen Census Divisions for the industrial group RETAIL. The upper and lower points of each distribution band correspond, respectively, to the 90:th and 10:th percentile of the distribution of the 500 values of $(\hat{t}_d - t_d)/t_d$. Consequently, a distribution band placed roughly symmetrically about the zero line indicates that the corresponding estimator is approximately unbiased for the domain of interest; otherwise, the estimator is biased for the domain. The shorter the band, the smaller the variance of the estimator in the domain. The abscissa measures the mean sample take for the domain.

From the tables and graphs, the following conclusions emerge: (where conclusion C states the main new results, whereas A and B resume what is known from earlier work Särndal and Råbäck (1983); Hidiroglou et al. (1984)).

- A. The SYN/C and SYN/R estimators are badly biased in some domains, namely, in those domains where the underlying model fits poorly. However, they consistently have an attractively low variance, compared to the other alternatives. The Mean Squared Error of the two SYN estimators will consequently be very large in domains with large bias (poor model fit); by contrast, the Mean Squared Error is small in domains with little bias (good model fit).
- B. The REG/C and REG/R estimators are essentially unbiased. Their variance, although usually much lower than that of the EXP and POS estimators, is consistently much higher than that of the SYN/C and SYN/R estimators. In the smallest domains, none of the unbiased estimators (EXP, POS, REG/C, REG/R) is attractive from the variance point of view; this is especially true for the REG estimators. This problem is remedied by the two MRE modifications of the REG estimators.
- C. The two MRE estimators, MRE/C and MRE/R, are negligibly biased when the SYN estimators happen to be nearly unbiased (e.g., RETAIL, area 17); otherwise the MRE estimators have a certain bias, which, however, is ordinarily much less pronounced than that of the SYN estimators (e.g., RETAIL, area 2). The MRE estimators have considerably smaller variance and Mean Squared Error, in all domains, than the REG estimators. This tendency is particularly pronounced in the smaller domains. In comparison with the SYN estimators, we find that the MRE estimators (as expected) still have a larger variance in virtually all domains. However, the Mean Squared Error of the MRE estimators is smaller than that of the SYN estimators in domains where the latter are badly biased. In Table 6 we see, for example, that the MRE/R estimator has a smaller Mean Squared Error than that of the SYN/R in 9 out of 16 small areas. The obvious explanation is that in domains where the SYN estimator is greatly biased, the $(\text{bias})^2$ constitutes an extremely large contribution to the Mean Squared Error of the SYN, whereas for the MRE estimators, the $(\text{bias})^2$ is not very important. Since we do not know which domains create the large biases, the goal of producing reliable estimates in all domains is on the whole better served by the MRE method of estimation.

Table 1
Mean Sample Take and Relative Bias of Each of Eight Estimators over
500 Repeated Simple Random Samples from the Entire Population
Industrial Group: RETAIL; 18 Census Divisions in Nova Scotia.

Area	Mean Sample Take	Estimator							
		EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	1.76	-0.02	-0.13	0.12	0.02	-0.03	0.30	0.09	-0.02
2	5.45	0.00	-0.04	-0.36	-0.10	-0.02	-0.27	-0.08	-0.02
3	3.90	-0.02	0.01	-0.08	-0.02	0.00	-0.01	-0.01	0.00
4	3.02	0.01	-0.05	0.15	0.05	0.01	0.13	0.04	0.04
5	5.93	0.00	0.01	0.21	0.05	0.00	0.13	0.03	0.00
6	7.63	-0.02	-0.01	0.28	0.07	0.01	0.10	0.02	0.00
7	8.61	0.02	0.01	-0.16	-0.03	0.01	-0.18	-0.03	0.01
8	5.64	-0.02	-0.01	0.34	0.10	0.03	0.24	0.06	0.01
9	24.64	0.00	0.00	-0.02	0.00	0.00	-0.01	0.00	0.01
10	8.92	-0.02	-0.02	0.15	0.02	-0.01	0.09	0.00	-0.01
11	8.35	-0.03	-0.02	0.08	0.01	0.00	0.10	0.02	0.00
12	10.58	0.01	0.00	-0.27	-0.05	0.00	-0.18	-0.03	0.00
13	0.48	-0.04	-0.58	0.61	0.36	0.04	1.00	0.58	0.04
14	2.80	0.03	-0.03	0.33	0.11	0.00	0.24	0.10	0.02
15	4.21	0.06	-0.01	0.28	0.06	0.00	0.30	0.07	-0.01
16	2.24	0.03	-0.05	0.74	0.26	0.03	0.94	0.32	0.02
17	23.95	-0.01	-0.01	-0.02	0.00	0.00	-0.05	-0.01	0.00
18	0.54	0.07	-0.54	0.63	0.34	-0.06	0.67	0.35	-0.06

Table 2
Mean Squared Error of Each of Eight Estimators over 500 Repeated Simple
Random Samples from the Entire Population
Industrial Group: RETAIL; 18 Census Divisions in Nova Scotia.

Area	Estimator							
	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	3,209	2,206	96	697	1,397	462	769	1,484
2	42,598	24,623	21,782	12,725	17,358	13,110	10,256	14,380
3	10,469	6,853	357	2,592	4,212	146	2,333	3,782
4	5,626	3,657	324	746	1,186	257	1,206	1,853
5	14,554	9,681	2,999	5,090	7,360	1,294	3,993	5,974
6	12,308	5,686	6,713	3,423	4,289	1,255	1,747	2,515
7	34,865	17,988	6,912	9,387	13,451	8,161	12,019	17,239
8	12,066	8,630	5,772	3,694	5,045	2,981	3,528	4,986
9	72,974	40,440	5,776	24,025	29,250	5,068	21,292	25,832
10	22,091	9,433	4,559	5,832	7,927	2,009	5,365	7,272
11	23,519	12,505	1,778	6,738	9,578	2,348	7,890	11,063
12	46,588	21,874	35,310	13,558	17,084	17,454	12,222	16,514
13	635	244	161	95	228	422	287	783
14	3,871	2,849	692	1,254	2,141	378	1,373	2,346
15	8,088	3,511	2,249	1,892	2,806	2,651	1,985	2,937
16	3,245	2,127	3,316	1,563	2,516	5,333	1,741	2,654
17	81,211	47,753	5,503	28,957	35,232	7,681	27,457	33,136
18	1,003	306	169	187	654	186	184	637

Table 3

Mean Sample Take and Relative Bias of Each of Eight Estimators over
500 Repeated Samples from the Entire Population
Industrial group: ACCOMMODATION; Areas: 16 Census Divisions in Nova Scotia.

Area	Mean Sample Take	Estimator							
		EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	0.25	0.01	-0.75	-0.08	-0.06	-0.01	0.36	0.28	0.01
2	1.37	-0.06	-0.21	0.25	0.10	0.02	0.25	0.11	0.02
3	1.02	0.06	-0.26	0.19	0.09	0.04	0.12	0.06	0.03
4	0.23	-0.10	-0.77	-0.33	-0.26	-0.07	-0.15	-0.13	-0.05
5	2.04	0.03	-0.13	0.21	0.08	0.03	0.18	0.06	0.01
6	1.49	0.04	-0.13	0.17	0.10	0.03	0.03	0.02	0.01
7	1.53	0.01	-0.18	-0.29	-0.11	-0.01	-0.30	-0.12	-0.02
8	1.54	0.03	-0.19	-0.42	-0.17	-0.01	-0.26	-0.11	-0.02
9	6.83	0.01	-0.02	0.13	0.02	0.00	0.12	0.02	0.00
10	1.26	-0.01	-0.26	0.40	0.17	0.03	0.30	0.13	0.02
11	3.06	0.04	-0.02	0.51	0.21	0.08	0.40	0.16	0.06
12	1.80	0.02	-0.16	-0.08	-0.05	-0.03	-0.23	-0.10	-0.03
14	1.04	0.02	-0.33	-0.52	-0.23	-0.07	-0.32	-0.15	-0.06
15	1.54	-0.03	-0.23	-0.21	-0.13	-0.08	-0.15	-0.11	-0.08
17	3.08	-0.07	-0.05	-0.03	-0.01	0.00	-0.14	-0.07	-0.03
18	0.52	0.04	-0.54	3.26	3.20	0.60	2.97	2.92	0.50

Table 4

Mean Squared Error of Each of Eight Estimators over 500 Repeated Simple
Random Samples from the Entire Population
Industrial Group: ACCOMMODATION; Areas: 16 Census Divisions in Nova Scotia.

Area	Estimator							
	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	1,142	283	9	7	25	58	44	164
2	7,467	5,082	877	631	1,077	747	455	726
3	878	442	48	163	242	24	116	163
4	155	43	7	6	17	3	3	6
5	15,200	8,392	2,091	2,270	3,230	1,271	1,208	1,785
6	5,239	3,906	253	1,038	2,193	54	396	792
7	21,197	8,781	3,569	1,831	3,016	3,709	1,812	2,948
8	14,071	6,738	3,608	2,122	4,018	1,492	947	1,766
9	50,606	27,867	9,980	11,413	14,344	6,575	7,779	9,991
10	2,219	993	590	362	665	317	151	280
11	10,535	5,774	6,366	5,126	7,154	3,867	2,752	3,673
12	16,787	10,485	543	1,148	1,944	1,245	1,130	1,836
14	51,471	25,644	9,669	8,221	14,155	3,972	3,189	5,077
15	59,207	41,381	4,861	10,548	18,119	2,759	4,262	6,636
17	29,632	25,211	1,501	3,023	4,754	1,765	2,123	3,214
18	286	99	2,062	2,112	5,623	1,607	1,646	4,561

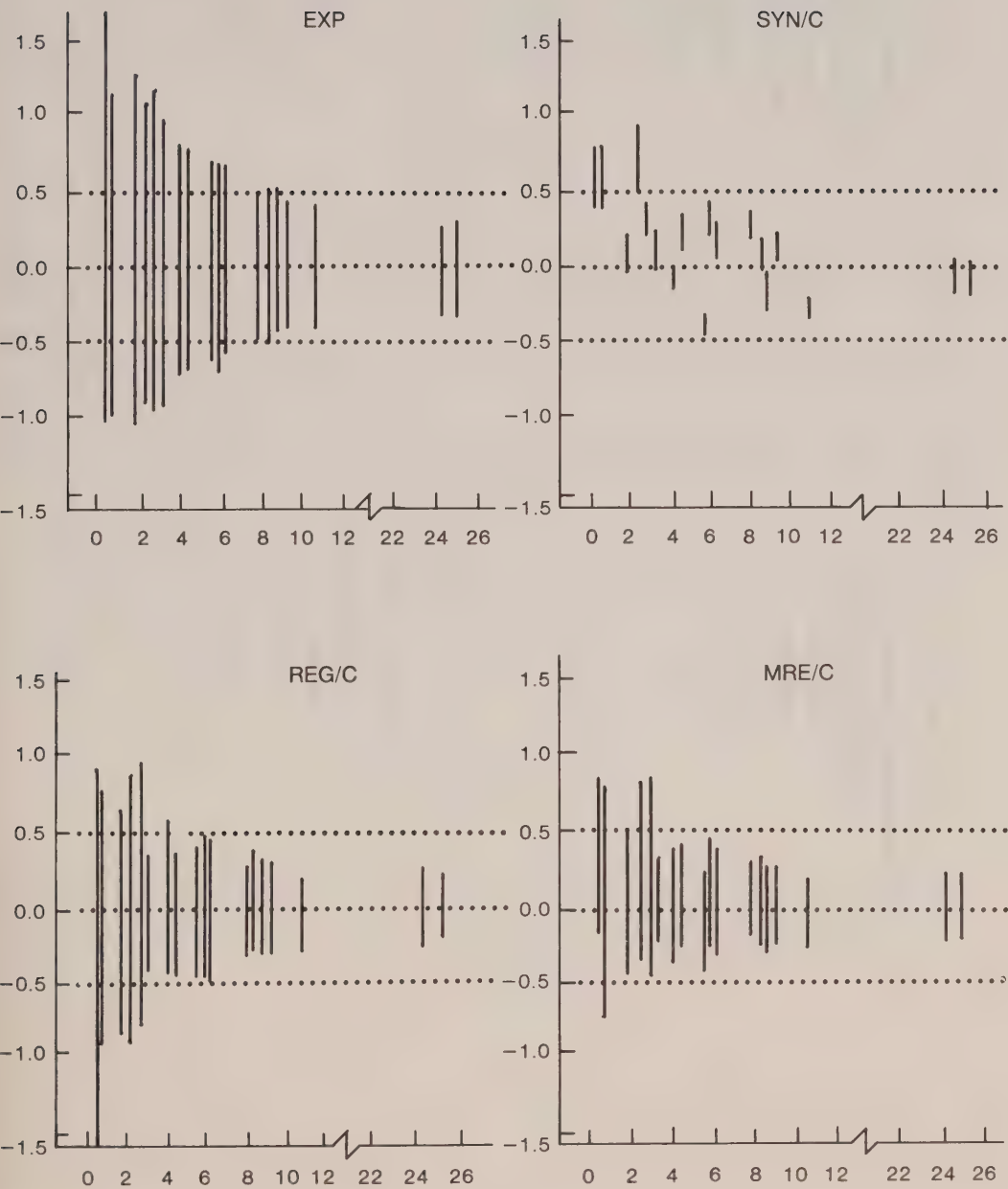
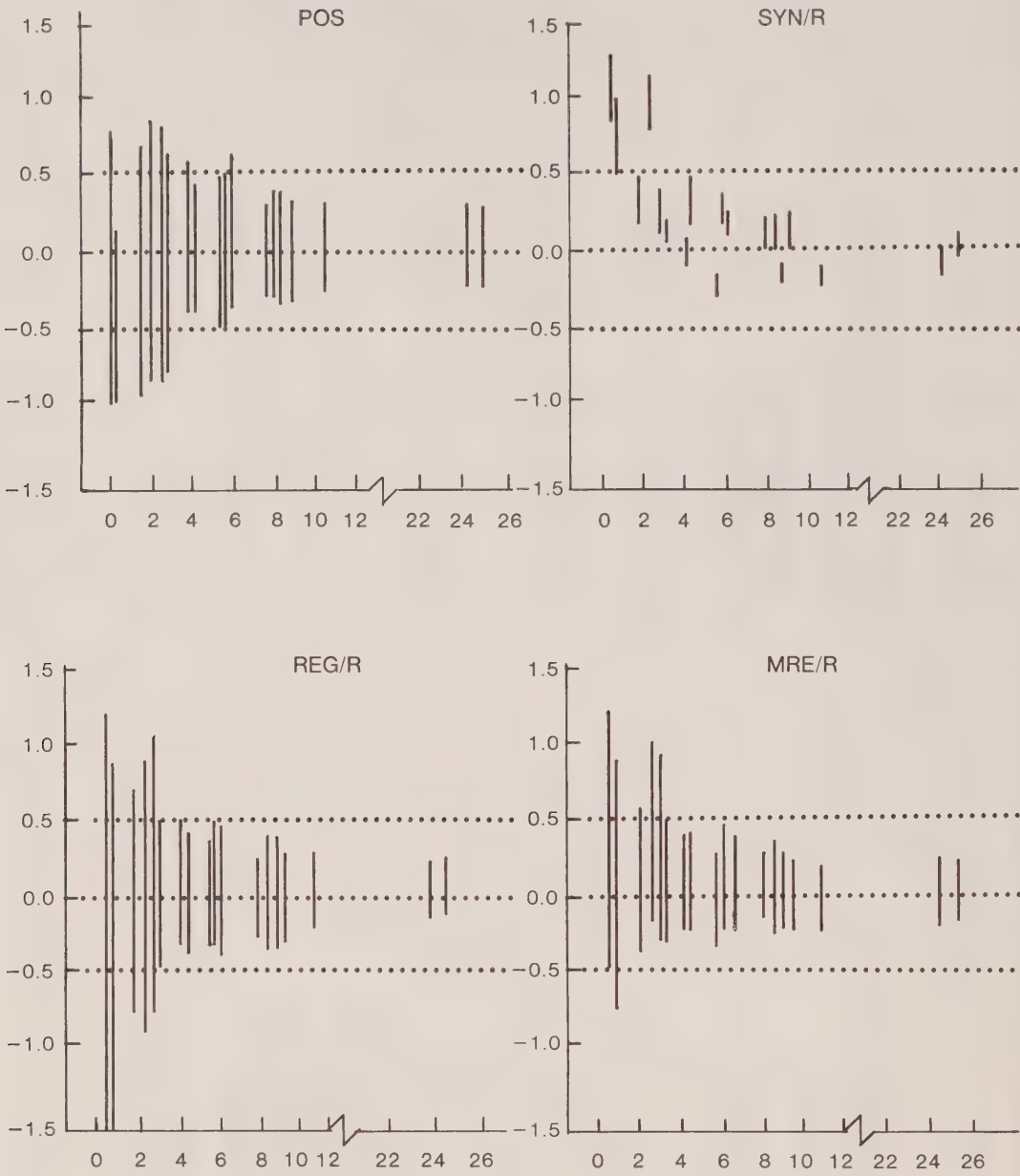


Figure 1: Distribution band of relative error for selected estimators — abscissa represents mean sample take. Industrial Group: RETAIL. Areas: 18 Census Divisions in Nova Scotia.

Figure 1 (continued)



5. CONCLUSIONS

In summary we find that the overall performance of the MRE estimators is such that we suggest them as promising alternatives for future applications of small area estimation. The recommended confidence interval procedure based on the MRE estimators is given in section 3.

We think that the MRE method presented here involves a simple mechanism for steering the estimates slightly in the direction of the stable SYN estimators, when the sample take is less than expected. This goal is also manifested (but attained by different means) in such other attempts as the empirical Bayes (Fay and Herriot, 1979) and sample-dependent (Drew, Singh, and Choudhry 1982) methods of estimation.

REFERENCES

- DREW, J.D., SINGH, M.P. and CHOUDHRY, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 17-47.
- FAY, R.E. and HERRIOT, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- HIDIROGLOU, M.A., MORRY, M., DAGUM, E.B., RAO, J.N.K. and SÄRNDAL, C.E. (1984). Evaluation of alternative small area estimators using administrative data. Paper presented at ASA meetings, Philadelphia, August, 1984.
- SÄRNDAL, C.E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustments for nonresponse. *Bulletin of the International Statistical Institute*, 49:1, 494-513. (proceedings, 43rd session, Buenos Aires).
- SÄRNDAL, C.E. and RÅBÄCK, G. (1983). Variance reduction and unbiasedness for small domain estimators. *Statistical Review*, 1983:5 (Essays in honour of T.E. Dalenius), 33-40.
- SÄRNDAL, C.E. (1984). Design-Consistent versus Model-Dependent Estimation for Small Domains. *Journal of the American Statistical Association*, 79, 624-631.

1981 Census of Agriculture Data Processing Methodology

DAVID K. HOLLINS¹

ABSTRACT

This paper presents an overview of the methodology used in the processing of the 1981 Census of Agriculture data. The edit and imputation techniques are stressed, with emphasis on the multivariate search algorithm. A brief evaluation of the system's performance is given.

KEY WORDS: Edit and imputation; Multivariable searches

1. INTRODUCTION

This paper presents an overview of the methodology used in the processing of the 1981 Census of Agriculture data. There are 3 separate phases to the processing of the data: Data Entry, Edit, and Imputation, each of which performs a different function. First, in Data Entry, data on the questionnaires are keyed onto a computer data file. Then, in the Edit phase, computer edits are applied to the keyed data records in order to detect any inconsistent, missing, or suspicious entries. In the final phase, Imputation, actions are taken to adjust the data records so that they conform to the rules defined by the computer edits applied during Edit. The methodology involved in each of the three phases of processing is described in subsequent sections of this paper. A flow chart of the 1981 Census of Agriculture processing is given in Figure 1.

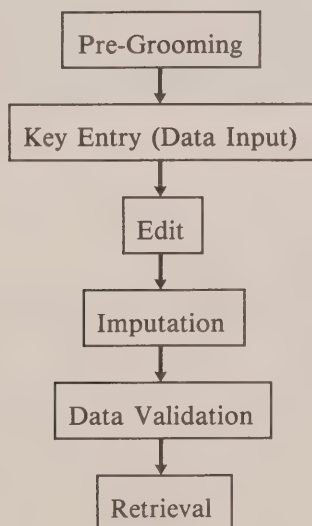


Figure 1. Overall Process Flow

¹ D.K. Hollins, Census and Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

The 1981 Census of Agriculture required that the same questionnaire be completed by each farm operator in Canada. The questionnaire is 8 pages long and consists of 134 questions. Questions are asked on all aspects of farm operation, including items such as types of crops grown, livestock raised, equipment maintained, and types of land use. Operators are required to answer only those sections of the questionnaire which apply to their holding.

As this paper is an overview, it is not possible to delve into the technical computer aspects of the Census of Agriculture processing. These details may be found in Shields and Yiptong (1981), on which this paper is based.

2. DATA ENTRY

In the Data Entry phase the Census of Agriculture data are transferred from the original questionnaires to a data file in computer memory. Data entry is comprised of two stages: a clerical pre-grooming process (Pre-Scan), and Key Entry.

After the questionnaires arrive at head office for processing, a clerical pre-grooming process known as Pre-Scan is performed. In this process, a clerk scans each questionnaire for response irregularities such as unreadable entries, ditto marks, and responses in incorrect locations. If valid responses can be discerned, they are recorded in the appropriate locations, if not, the questionnaire is left unchanged.

Next, in Key Entry, the data on each questionnaire are keyed into the computer. Identifying information from the front page of the questionnaire is entered in a standard fixed format. However, since farm operators are required to answer only the sections of the questionnaire that apply to their holding, a large portion of the questionnaire remains blank. To reduce keying time, a method known as "string-keying" is used to enter the remaining data. This means that the field name is keyed, immediately followed by the data value for that field. Only fields with existing data values are keyed; unanswered portions of the questionnaire are not. Because of the sparseness of the data, this method results in significant savings in keying time required.

The Key Entry process creates one Edit and Imputation Master File (EIMF) record for each of a total of approximately 320,000 questionnaires. There are 244 fields on an EIMF record, each identified by a name, generally 6 characters in length. The Key Entry operator is instructed to key "#" for any unreadable entries. If possible, a clerical correction will be performed on records containing this symbol during Edit, otherwise, the records will be corrected during imputation.

3. EDIT

The Edit phase serves two purposes. The first is to use computer edits to detect any inconsistent, missing, or suspicious entries in the data. The second is to perform a clerical correction on the defective records, or if that is not possible, then to pass the defective records on to be fixed during Imputation. A flow chart of the Edit process is given in Figure 2.

There are 3 components to the edit system: two computer edit cycles called Correction Cycles #1 and #2, and a cycle for correcting edit failures, called Correction of Rejects. Correction Cycle #1 (CC #1) consists of those edits that detect conditions that prevent the "de-stringing" (the conversion from string format to fixed format) of the keyed record (decode edits), and those edits that detect errors in the geographic and identifying information from the front page of the questionnaire (ID edits). Correction Cycle #2 (CC #2) consists of those edits that identify inconsistencies in the main body of the data (data edits). Correction of Rejects is a clerical process during which both CC #1 and CC #2 edit failures are corrected manually. Edit failures that cannot be corrected by Correction of Rejects are passed on to Imputation.

Each of the EIMF records is processed through the edit system individually.

3.1 Correction Cycle #1 (Decode and ID Edits)

Correction Cycle #1 consists of the application and resolution of two sets of edits: the decode edits and the ID edits.

The decode edits are applied first and if conditions exist that prevent the “de-stringing” of the data record, then decode edit failures will result. For example, as no two fields should have the same identifying characters, “de-stringing” will be prevented if two field names are keyed identically.

Any failed decode edits are resolved manually by the Correction of Rejects staff. This involves returning to the questionnaire to determine the cause of the edit failure, then the rekeying of the relevant data. After an attempt is made to resolve a decode edit failure, the EIMF record is re-edited by passing it through the decode edits again, forming a continuous cycle between the decode edits and the Correction of Rejects staff. This cycle is repeated until there

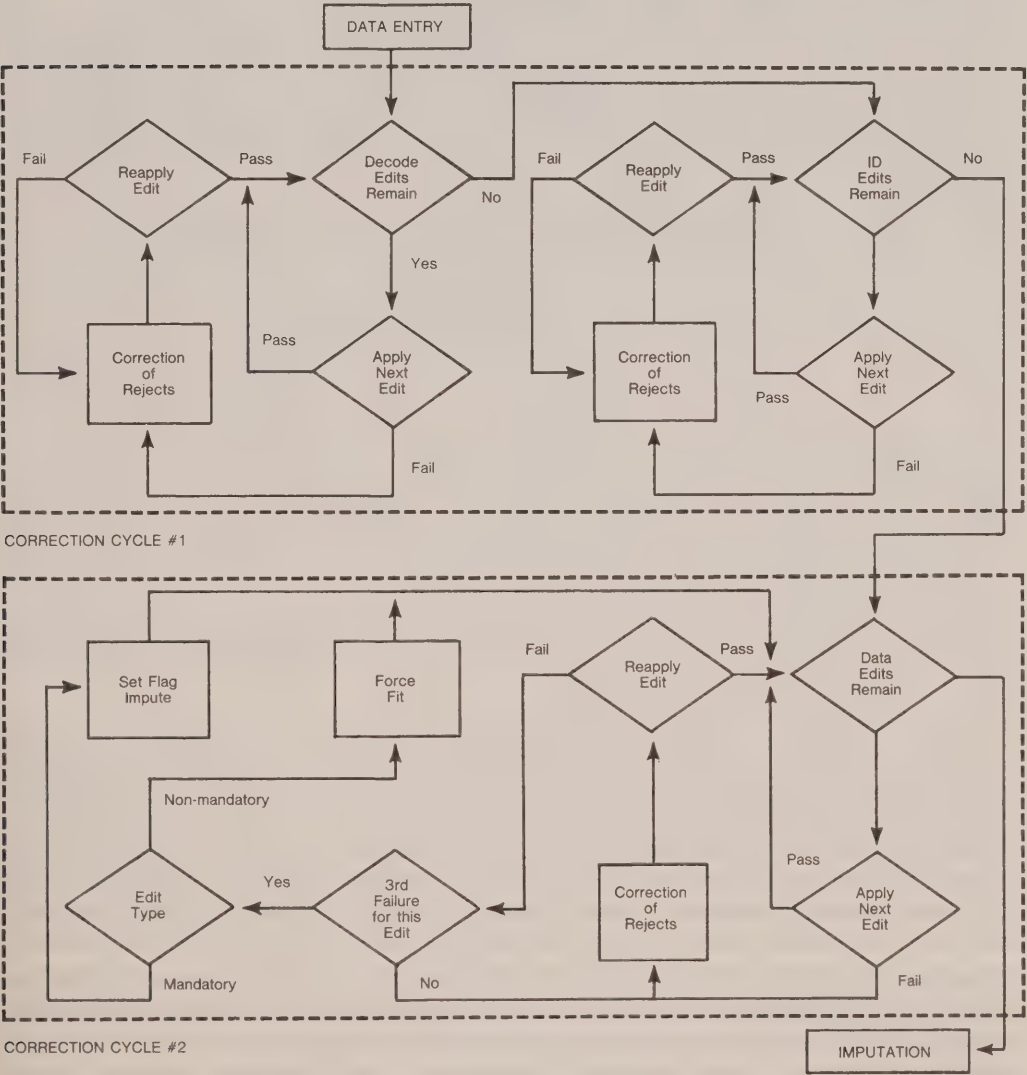


Figure 2. Edit Process Flow

are no decode edit failures remaining on the EIMF record. If a decode edit cannot be resolved directly, the most appropriate valid interpretation of the available data is employed as a final override.

After all decode edit failures have been resolved, the ID edits are applied. If any of the identifying information on the EIMF record is inconsistent or missing, then one or more ID edits will fail. These ID edit failures are resolved in an identical manner to the decode edits.

Once all of the CC #1 (decode and ID) edit failures have been resolved by the Correction of Rejects staff, the EIMF record is passed through the CC #2 edit program.

3.2 Correction Cycle #2 (Data Edits)

The data edits (CC #2) are used to detect errors in the main body of the questionnaire, as opposed to errors in coding, or in identifying information. There are two types of data edits: non-mandatory edits (75), and mandatory edits (24).

Non-mandatory edits are written to detect suspicious entries on the EIMF data records. Generally, non-mandatory edits, detecting variable values falling outside prescribed limits, are performed by comparing different fields or groups of fields on the questionnaire to determine if some data values are abnormally high or low in comparison with others. For example, a record with total farm area equalling 10 acres and containing 10,000 cattle would be flagged by a non-mandatory limit edit.

Mandatory edits are written to detect logical impossibilities on the data record, e.g., if the total number of cattle reported is not equal to the sum of the reported values for each of the different cattle types, then a mandatory edit would fail. The most complex mandatory edits are those written for the crop section of the questionnaire.

To resolve a non-mandatory edit failure, the record is sent to a Correction of Rejects clerk. The Correction of Rejects clerk first notes whether or not the edit failure is due to a keying error. If it is, the relevant data is rekeyed. If it is not, the clerk scans the questionnaire to see if the respondent has written any comments on the questionnaire that may explain the reason for the edit failure. For example, if the respondent is instructed to answer a question in tons, and tons has been crossed out and pounds written in, the response will probably fail a non-mandatory limit edit. In this case, the Correction of Rejects clerk will convert the response from pounds into tons. If the Correction of Rejects clerk can find no explanation for the edit failure, the respondent's answers are left intact on the EIMF record and are indicated acceptable. Although no changes are made to the data on the EIMF record, this is known as "force-fitting" the data.

Mandatory edit failures are handled somewhat differently to non-mandatory edit failures. To resolve a mandatory edit failure, the failed record is sent to a Correction of Rejects clerk who proceeds at first in an identical manner to that used in the resolution of non-mandatory edit failures. However, if no explanation for the edit failure can be found, instead of "force-fitting" the edit failure, the record is flagged for computer imputation.

As in CC #1, there is a continuous cycle between the Correction of Rejects staff and the CC #2 edit program. After each attempt is made to resolve a CC #2 edit failure the EIMF record is re-run through the CC #2 edit program. Unlike CC #1, however, the Correction of Rejects clerk has only 3 attempts to resolve the CC #2 edit failures on a given EIMF record. After the third attempt, the CC #2 edit program is run once again. Any remaining non-mandatory edit failures are marked "force fit" and any remaining mandatory edit failures are marked "impute". The mandatory edit failures are simply flagged at this stage. The particular fields requiring imputation are identified at the imputation stage.

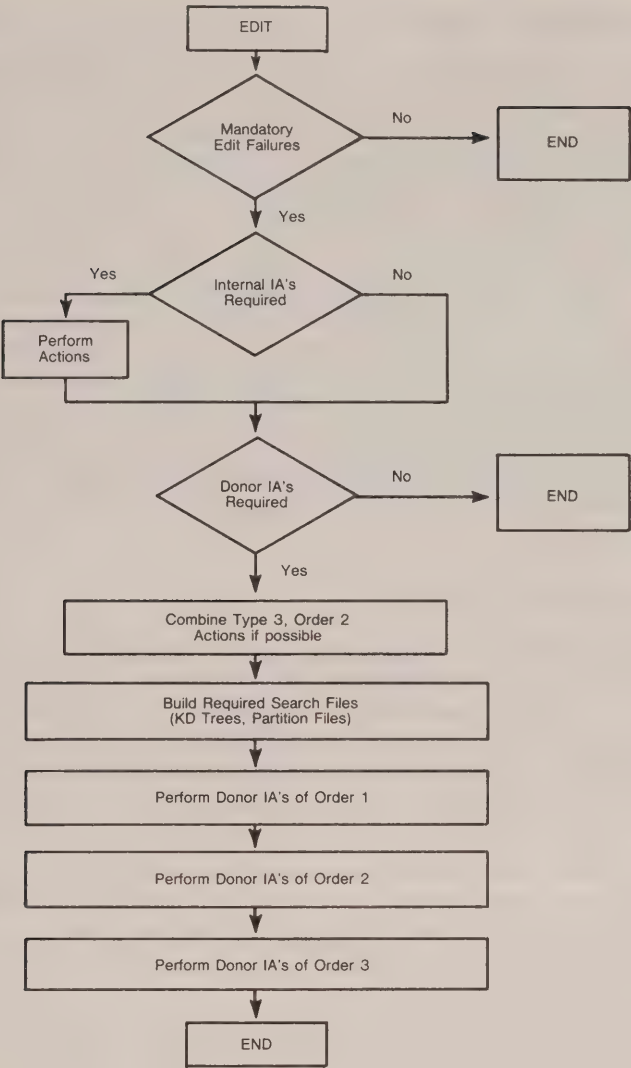


Figure 3. Imputation Process Flow

4. IMPUTATION

The purpose of the 1981 Census of Agriculture imputation system (see Figure 3) is to resolve edit failures on the EIMF data records. As all non-mandatory edit failures are “force-fit” as described in the previous section, only the mandatory edit failures remain to be resolved by the imputation system. In order to make the EIMF data records conform to the mandatory edits, specified “imputation actions” are performed. These imputation actions (IA’s), of which there are over 100, are designed so that as few fields as possible are changed on the EIMF record, e.g. totals are always adjusted to equal the sum of the parts, rather than the parts being adjusted to total the sum. Each IA has associated with it the appropriate imputation processing control information and is selected based on the field or fields requiring imputation. There are two different types of IA’s performed: internal IA’s, or deterministic corrections, and donor IA’s.

4.1 Internal Imputation Actions

Internal IA's are performed in cases where sufficient data exists on the failed record to enable the imputation system to provide a deterministic correction for the inconsistent field(s). These internal IA's are performed in cases where the inconsistent field(s) is (are) deterministically dependent on other fields not requiring imputation. For example, an internal IA would be performed if a respondent reports quantities for the various types of cattle but neglects to report the total number of cattle. In this case, total cattle would be calculated using the sum of the quantities reported for the various types of cattle. Another situation in which an internal IA would be performed is where a respondent reports a certain quantity of a particular type of fruit tree but neglects to give the corresponding acreage. In this case, the acreage would be computed using a predetermined average density for that type of fruit tree. Internal IA's are performed in accordance with constraints to ensure that the imputed values are within reasonable bounds.

The implementation of internal IA's is more straightforward than that of donor IA's. As the internal IA is performed using data from the same record, there is no need to specify an algorithm for donor selection. The only requirement is to perform the deterministic correction specified by the appropriate internal IA. All internal IA's are performed before proceeding to donor imputation.

4.2 Donor Imputation Actions

When the inconsistent field or fields are not deterministically dependent on other consistent fields, internal IA's cannot be applied. The lack of sufficient information on the failed record to provide a deterministic correction to the inconsistent field(s) necessitates an imputation method using data contained on another record. This method, known as donor imputation, involves the transfer of data from a "clean" donor record (one which has passed all mandatory edits) to the failed record. The transferred data will restore consistency to the inconsistent field(s) on the failed record. For example, a donor IA will be performed in order to estimate the distribution for types of cattle when only the total number of cattle is reported. In this case, the distribution of cattle types present on the donor record is transferred to the failed (recipient) record.

As donor imputation requires an algorithm for locating a donor record, it is more complex to implement than internal imputation. In order to perform donor imputation, several search "parameters" must be specified.

To ensure that a "clean" donor record is geographically close to the "bad" recipient record, the country is divided into distinct geographical regions called imputation regions. The delineation of these imputation regions is based on the existing "crop district" boundaries which are defined according to characteristics such as soil type and climate. There are 59 crop districts, and thus 59 imputation regions, in Canada with an average of 5,500 farms per region. In order to be an eligible donor, a record must be in the same imputation region as the recipient record.

In order to avoid searching records that cannot donate suitable data, each donor IA also specifies the subpopulation on which the donor search is to take place. For example, if the distribution for types of cattle is being imputed, then the only records searched in order to find a donor would be members of the subpopulation where cattle have been reported. A given record may be a member of several of the 30 different subpopulations. In some cases, all clean records within the imputation region are deemed suitable donors in which case the general population in the imputation region is defined as the appropriate subpopulation.

The final constraint on the file of eligible donors is the fact that records requiring any donor imputation themselves cannot be used as donors. However, records requiring only internal imputation may be used as donors.

In summary, the file of eligible donors consists of all records not requiring donor imputation that are members of the subpopulation specified by the imputation action to be performed and that are also located in the same imputation region as the bad record.

As some records require more than one IA to be performed, there is need for a hierarchical system of imputation action execution. To specify the order in which the IA's are to be performed, every IA, both internal and donor, has one of three "orders" associated with it. IA's of order 1 are performed first, followed by IA's of orders 2 and 3 respectively.

To aid in the selection of a suitable donor record, one or more variables not requiring imputation are selected to be used as matching variables for each donor IA. These matching variables, selected by subject matter experts, are considered to be highly correlated with the field(s) requiring imputation. Both the recipient and the selected donor record should have similar matching variable values. As the use of continuous matching variables does not permit exact matches, a distance function based on the selected matching variable(s) is used to identify the closest eligible donor to the bad record.

Each donor IA has one of three possible search types associated with it. Partition searches (type 1) are performed when only 1 discrete matching variable is specified for the IA. Binary searches (type 2) are performed when only 1 continuous matching variable is specified for the IA. Multivariable searches (type 3) are performed when 2 or more continuous matching variables are specified for the IA. Each of these three search types is described individually in the following sections. Other combinations of matching variable types are not employed.

Finally, after a suitable donor has been selected and if specified in the IA control information, the donated data from the donor record are prorated before transferring them to the recipient record. For example, if the variable "number of trucks" is used as a matching variable for imputing "value of trucks", then the value of "value of trucks" assigned to the recipient record is equal to "value of trucks" of the donor, multiplied by the ratio "number of trucks" of the recipient divided by "number of trucks" of the donor.

As previously described, each donor imputation action has one of three search types associated with it. Two of these search types, binary and partition searches, are used to perform imputation actions for which only 1 matching variable is specified. The other search type, the multivariable search, is performed when 2 or more continuous matching variables are to be used.

4.2.1 Type 1 — Partition Searches

Partition Searches are performed when only 1 discrete matching variable with a small number of possible values is specified for the imputation action, e.g., as in the case where a respondent reports the total number of tractors, but neglects to give the corresponding total dollar value. Since a farmer is unlikely to have more than 3 tractors the donor population is divided into 3 partitions: 1, 2, or 3+ tractors. A donor is chosen at random from the partition to which the recipient record belongs. If there are no donor records within the partition to which the recipient record belongs, but there are donors in any of the subsequent (higher numbered) partitions, then all of the subsequent partitions are collapsed into one and a donor record is selected at random from this collapsed partition. If there are no donor records in the partition to which the recipient record belongs or in any subsequent partition, then a donor record is selected at random from the closest preceding (lower numbered) partition that contains any donor records. As these collapsing procedures are not frequently applied, no serious introduction of bias is encountered. If the donor population is empty, then the field to be imputed is assigned the maximum value allowable by the edits and the record flagged to indicate that imputation was unsuccessful. These flagged records are then reviewed by subject matter personnel who manually assign an appropriate value to the field requiring imputation.

4.2.2 Type 2 — Binary Searches

Binary searches are performed when only 1 continuous matching variable is specified for the imputation action, e.g., as in the case where a respondent reports the total value of his/her tractors, but does not give the corresponding number of machines. The entire file of eligible

donor records is searched and the record that minimizes the difference between the matching variable values is selected as the donor. If two or more potential donor records are equally close, then the one that is geographically closer to the recipient (as judged from the geographic ID) is automatically selected as the donor. If the donor population is empty, then the recipient record is flagged to indicate that imputation was unsuccessful.

4.2.3 Type 3 — Multivariable Searches

Multivariable searches are performed when more than one continuous matching variable are specified for the imputation action. These are the most complex of the three search types performed by the 1981 Census of Agriculture. The method used to perform multivariable searches was adapted for use at Statistics Canada by G. Sande.

When the missing data are related to more than one continuous matching variable, it is desirable to use as a donor a record that is closest to the recipient record on all these matching variables simultaneously. This requires a multivariable search on a large donor file and has been made practical by grouping the donor population in such a way that it is not necessary to search every donor to determine the closest. This specialized grouping of records is called the K-D (Key Discriminator) tree. The same K-D tree may be used for all records requiring a certain donor IA within a particular imputation region as the file of eligible donors will remain the same in each case. However, if a different donor IA is to be performed using a different donor population, or even the same donor IA on a different imputation region, a new K-D tree must be built as the file of eligible donors will not contain the same records.

a) Building the K-D Tree

The first step in the building of the K-D tree is to perform a transformation on all of the matching variables by subtracting the mean and dividing by the standard deviation of the donor population. This allows matching variables of different scales to be specified for the same search.

After the variable transformation, the following algorithm is then used to actually build the K-D tree. It is first applied to the entire file of eligible donors, and then to all subfiles subsequently created by the algorithm.

Firstly, the range (largest value minus smallest value) is calculated for each of the matching variables specified. The median value of the variable with the largest range (or the variable with the smallest ID if there are 2 or more with the maximum range) is then calculated. The variable for which the median is calculated is called the discriminator variable. This median value is used to split the file into 2 new subfiles, the left subfile containing records with values less than or equal to the median value of the discriminator variable, and the right subfile containing records with values greater than the median value of the discriminator variable. The algorithm is then progressively re-applied to the resulting subfiles using all specified matching variables until all files become TERMINAL, at which point the building of the K-D tree is complete. A subfile becomes TERMINAL when either the range equals zero for all matching variables, i.e., all records in the subfile are identical, or if there are 16 or less records in the subfile.

The above algorithm will yield a K-D tree of the form illustrated in Figure 4.

Every record contained in the original file will be present in one and only one of the subfiles corresponding to the terminal nodes.

b) Searching the K-D Tree

In order to locate the best possible donor, it is necessary to decide which of the terminal nodes "corresponds" to the recipient record. This is done by traversing the K-D tree, using the transformed matching variable values of the recipient record, starting with the root node and proceeding until one of the terminal nodes is reached. At each node of the tree it is determined, using the discriminator variable for that node, which of the two lower nodes the recipient record corresponds to. The K-D tree is traversed in this manner until a terminal node is reached.

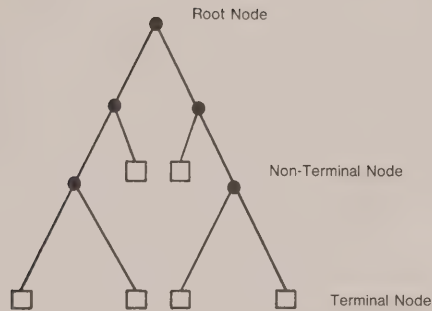


Figure 4. General Form of K-D Tree

In order to determine which donor in the chosen terminal node is closest to the recipient record, a distance function is required. Because of its ease of implementation, the distance defined by the maximum of the absolute differences between matching variables was used. The selected donor record is the one that minimizes this “distance”.

Although the selected donor record is the closest to the recipient record contained in the chosen terminal node, it is possible that there are closer donor records residing in other

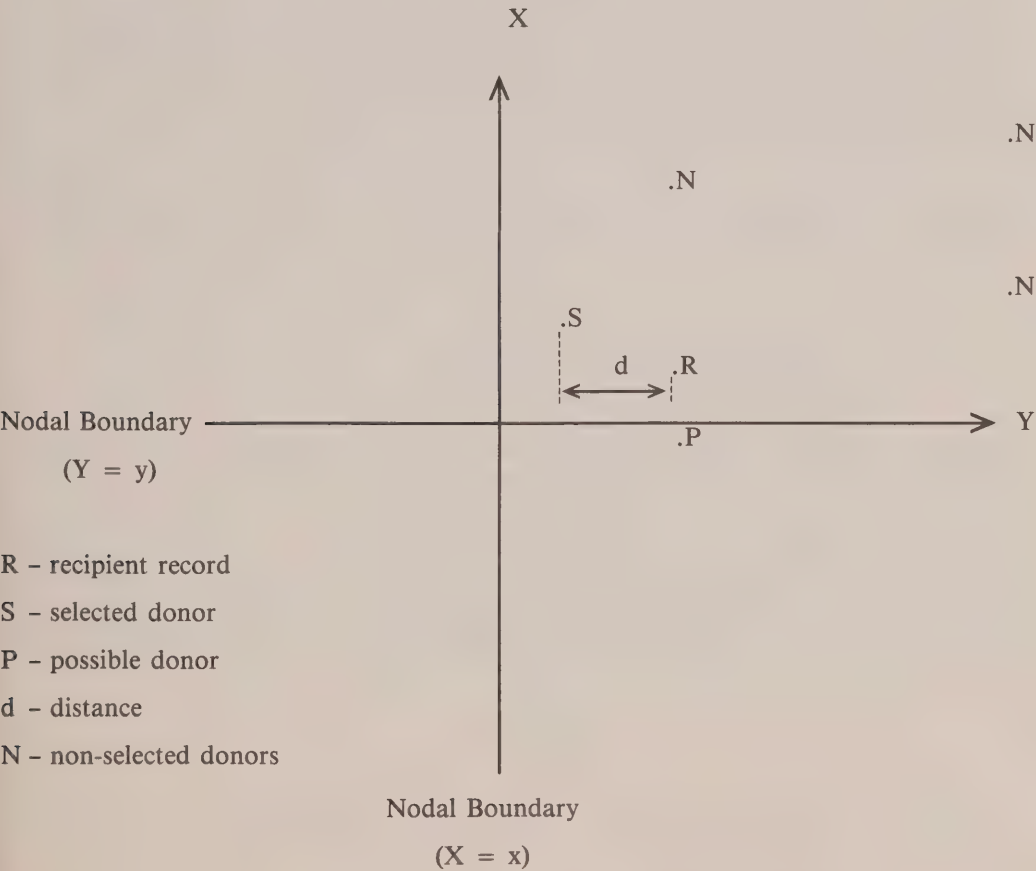


Figure 5. Closer Donors From Other Terminal Nodes (two matching variables)

terminal nodes. This may occur only if a nodal boundary exists that is closer to the recipient record than the currently selected donor record. This case is shown in Figure 5 for a donor IA involving two matching variables; X and Y. Each quadrant represents a terminal node.

It is evident that the possible donor P is closer to the recipient R than the selected donor S. This is possible because R is closer to the position of the nodal boundary $Y = y$ than to S, and only donor records lying in the same terminal node as the recipient record may be selected.

A procedure, based on the variable values used to define the nodal boundaries and known as the bounds-overlap-ball (B.O.B.) test, is used to determine which of the other terminal nodes, if any, may contain donors closer to the recipient record than the selected donor record. Only terminal nodes that have the potential to provide closer donors are tested, and if a closer donor is found, then it replaces the previously selected donor. The B.O.B. test is applied until all nodes that may contain closer donors have been tested.

Finally, for all three search types, after the eventual donor record has been selected, the donated data values are prorated as previously described, if specified in the IA control information.

It will always be possible to select a donor unless the donor population is empty. If this occurs then the imputation region is collapsed with another and imputation is redone. It was never necessary to perform this operation in 1981.

5. CONCLUDING NOTE

A detailed evaluation, Grenier (1983), indicated that a major portion of the edit system was of little data quality benefit. This was because the Correction of Rejects procedures were unable to correct a sufficient proportion of the edit failures. For example, Correction of Rejects was unable to correct the failures resulting from a subset of 77 of the 97 edits more than 5% of the time. Also, many of the edits affected less than .1% of the population. Additionally, the Correction of Rejects procedures were highly labour intensive and created a heavy paper burden. To eliminate these inefficiencies a new computer edit system will be designed for 1986.

Statistics from the 1981 Census of Agriculture, Grenier (1983), indicated that 43% of the farms in Canada had at least one field imputed. Of this 43%:

- 18% required internal imputation only,
- 17% required donor imputation only, and
- 8% required both internal and donor imputation.

An analysis of the data distributions before and after imputation indicated that the imputation system did not have a serious impact at the Canada level although many of the 137,390 records imputed underwent a significant change. The system successfully handled all necessary imputations with only 58 records requiring manual imputation. The system was found to be very efficient, a processing cost of only \$15,000 being incurred. Diagnostic data indicated that minor modifications to the system must be made for greenhouses, mushroom houses, community pastures, and institutions, if they are to remain in the census. Due to its successful fulfillment of the requirements, it is planned to reuse the present imputation system in 1986.

REFERENCES

- SHIELDS, M., and YIPTONG, J. (1981). Census of Agriculture-1981 Imputation Specifications. Technical Report, Census and Household Survey Methods Division, Statistics Canada.
- GRENIER, A.R. (1983). 1981 Census of Agriculture Evaluation Report. Technical Report, Agriculture Statistics Division, Statistics Canada.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8 1/2 par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1, I).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

BIBLIOGRAPHIE

- SHIELDS, M., et YIPTONG, J. (1981). Census of Agriculture-1981 Imputation Specifications. Rapport technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- GRENIER, A.R. (1983). 1981 Census of Agriculture Evaluation Report. Rapport technique, Division de la statistique agricole, Statistique Canada.

Une procédure faisant appel aux valeurs des variables pour définir les frontières nodales et appelée "bounds-overlap-ball test" (test B.O.B.) est utilisée pour trouver quels noeuds parmi les autres noeuds terminaux, s'il y en a, peuvent contenir des enregistrements dont-neurs situés plus près de l'enregistrement receveur que ne l'est l'enregistrement donneur choisi. Seuls les noeuds terminaux susceptibles de contenir des enregistrements donneurs plus proches sont testés et, advenant le cas où un enregistrement donneur plus proche est trouvé, il remplace l'enregistrement donneur choisi en premier lieu. Le test B.O.B. est répété jusqu'à ce que tous les noeuds susceptibles de contenir des enregistrements donneurs plus proches aient été testés.

Enfin, pour les trois types de recherche, une fois que l'enregistrement donneur éventuel a été choisi, les valeurs des données de l'enregistrement donneur dont on veut se servir pour faire l'imputation sont multipliées, si les renseignements de contrôle de la PI le prévoient, par un coefficient de proportionnalité comme on l'a indiqué précédemment.

Il sera toujours possible de choisir un enregistrement donneur à moins que la population des enregistrements donneurs ne soit vide. Dans ce cas, la région d'imputation est fondue dans une autre et l'imputation est refaite. Il n'a jamais été nécessaire d'exécuter cette opération pour les données du recensement de l'agriculture de 1981.

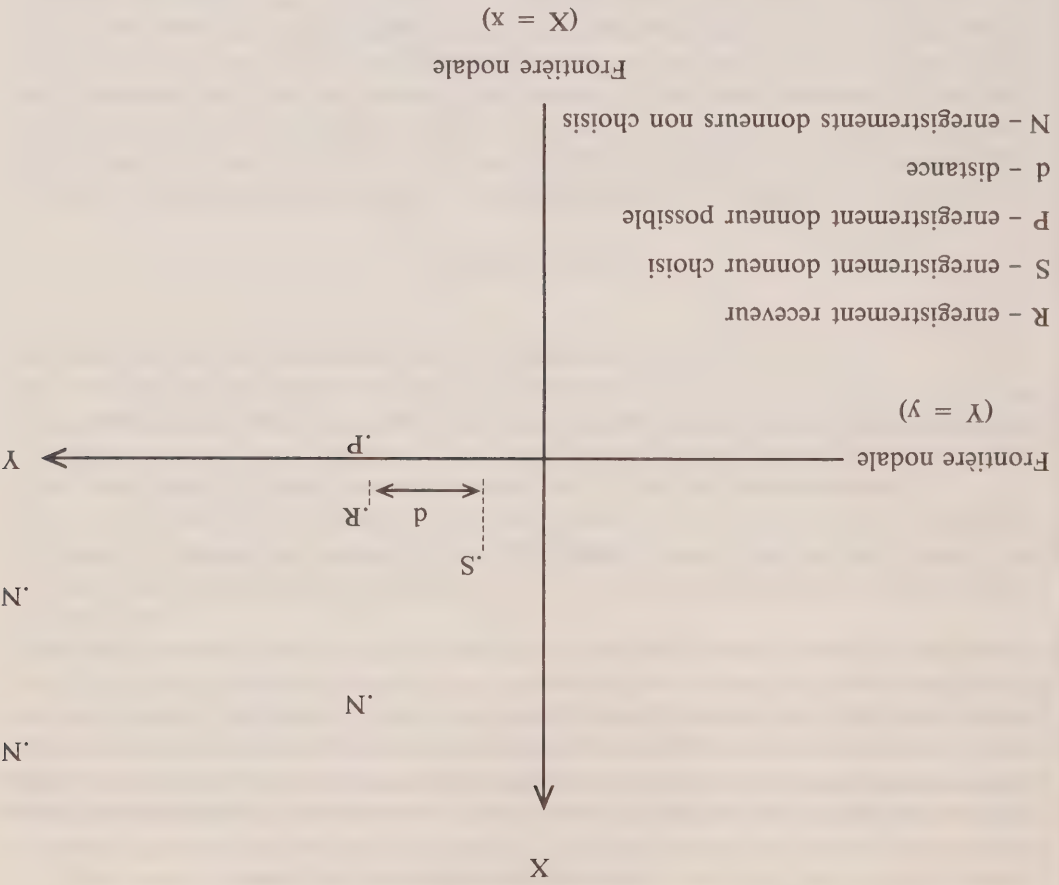
5. CONCLUSION

Dans une étude d'évaluation détaillée, Grenier (1983) indique que dans l'ensemble le système de contrôle a très peu contribué à améliorer la qualité des données. Cela est attribuable au fait que les procédures de correction des rejets n'ont pu corriger une proportion suffisamment grande des rejets sur vérifications. Par exemple, le personnel responsable de la correction des rejets n'a pas réussi à corriger dans plus de 5% des cas les enregistrements rejetés par un sous-ensemble de 77 vérifications sur 97. Également, bon nombre des vérifications touchaient moins de 0.1% de la population. De plus, les procédures de correction des rejets ont exigé beaucoup d'heures de travail et la manipulation d'une grande quantité de documents. Pour éliminer ces facteurs d'inefficacité, un nouveau système de contrôle informatisé sera conçu pour 1986.

Dans son analyse des statistiques du recensement de l'agriculture de 1981, Grenier (1983) indique que les enregistrements de 43% des exploitations agricoles au Canada ont au moins une zone dont la valeur est imputée. Plus précisément,

18% ont fait l'objet d'une imputation interne uniquement,
17% ont fait l'objet d'une imputation "d'emprunt" uniquement et
8% ont fait l'objet des deux types d'imputation.

Une analyse des différentes catégories de données avant et après imputation montre que le système d'imputation n'a pas eu beaucoup d'effet sur les résultats au niveau de l'ensemble du Canada même si bon nombre des 137,390 enregistrements ayant fait l'objet d'une imputation ont été sensiblement modifiés. Le système a permis d'effectuer toutes les imputations nécessaires, seulement 58 enregistrements ayant dû faire l'objet d'une imputation manuelle. Le système s'est avéré très efficace, il a coûté seulement \$15,000 à exploiter. Des données de diagnostic ont indiqué que des modifications mineures doivent être apportées au système dans le cas des serres, des champignons, des pâturages communautaires et des institutions pour que ces différentes catégories d'exploitations agricoles puissent être conservées dans le recensement. Comme le système d'imputation sous sa forme actuelle a bien répondu aux exigences, on prévoit le réutiliser en 1986.



a) Construction de l'arbre "K-D"

La première étape de la construction d'un arbre "K-D" consiste à exécuter une transformation sur toutes les variables d'appariement en soustrayant la moyenne et en divisant le résultat obtenu par l'écart-type de la population des enregistrements donneurs. Cela permet de spécifier les variables d'appariement d'échelles différentes pour la même recherche.

Une fois les variables transformées, on utilise l'algorithme qui suit pour la construction proprement dite de l'arbre "K-D". L'algorithme est d'abord appliqué à l'ensemble du fichier des enregistrements donneurs acceptables, puis à tous les sous-fichiers résultant de la première application de l'algorithme.

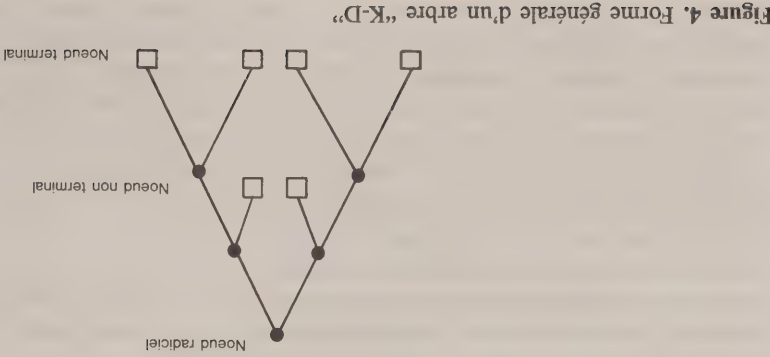
On commence par calculer l'étendue (la plus grande valeur moins la plus petite valeur) pour chacune des variables d'appariement spécifiées. On calcule ensuite la valeur médiane de la variable ayant la plus grande étendue (ou de la variable ayant le plus petit numéro de code d'identification s'il y a deux variables ou plus qui ont la même étendue et que celle-ci est la plus grande). La variable pour laquelle on calcule la médiane est appelée la variable de discrimination. La valeur médiane de cette variable sert à diviser le fichier en deux nouveaux sous-fichiers, le fichier de gauche contenant les enregistrements dont les valeurs sont inférieures ou égales à la valeur médiane de la variable de discrimination et le fichier de droite contenant les enregistrements dont les valeurs sont supérieures à la valeur médiane de la variable de discrimination. L'algorithme est ensuite progressivement réappliqué aux sous-fichiers résultants pour toutes les variables d'appariement spécifiées jusqu'à ce que tous les fichiers deviennent TERMINAL, étape où la construction de l'arbre "K-D" est achevée. Un sous-fichier devient TERMINAL quand l'étendue égale zéro pour toutes les variables d'appariement, c'est-à-dire lorsque tous les enregistrements du sous-fichier sont identiques, ou quand il y a 16 enregistrements ou moins dans le sous-fichier.

L'algorithme qu'on vient de décrire aboutira à la construction d'un arbre "K-D" de la forme illustrée à la figure 4.

Chacun des enregistrements contenus dans le fichier initial se retrouvera dans un seul des sous-fichiers correspondant aux noeuds terminaux.

b) Recherche d'un donneur

Pour situer le meilleur enregistrement donneur possible, il faut décider lequel des noeuds terminaux "correspond" à l'enregistrement receveur. Cela est fait en parcourant l'arbre "K-D" à l'aide des valeurs indiquées dans l'enregistrement receveur pour la variable d'appariement transformée, en commençant au noeud racine et en continuant jusqu'à ce qu'un des noeuds terminaux soit atteint. À chaque noeud de l'arbre, on peut savoir, à l'aide de la valeur de la variable de discrimination à ce noeud, auquel des deux noeuds inférieurs correspond l'enregistrement receveur. L'arbre "K-D" est parcouru de cette façon jusqu'à ce qu'un noeud terminal soit atteint.



à laquelle appartient l'enregistrement donneur. S'il n'y a aucun enregistrement donneur dans la partie à laquelle appartient l'enregistrement receveur, mais qu'il y a des enregistrements dans l'une ou l'autre des parties subséquentes (dans les parties constituées des enregistrements indiquant un plus grand nombre de tracteurs), alors toutes les parties subséquentes sont regroupées en une seule et l'enregistrement donneur est choisi au hasard dans la partie ainsi formée. S'il n'y a d'enregistrement donneur ni dans la partie à laquelle l'enregistrement receveur appartient, ni dans les parties subséquentes, l'enregistrement donneur est alors choisi au hasard dans la première partie à contenir des enregistrements donneurs et à précéder la partie à laquelle appartient l'enregistrement receveur (dans la première partie à contenir des enregistrements donneurs). Etant donné que l'on a très rarement recours à ces procédures de regroupement, aucun biais important n'est introduit dans les résultats. Si la population des enregistrements donneurs est vide, on attribue à la zone dont la valeur doit être imputée la valeur maximum permise par les vérifications et l'enregistrement est marqué d'un signe indiquant que l'imputation n'a pas réussi. Les enregistrements ainsi marqués sont ensuite examinés par le personnel concerné qui attribue manuellement une valeur à la zone dont la valeur devait être imputée.

4.2.2 Type 2 — Recherches dichotomiques

Les recherches dichotomiques sont effectuées quand une seule variable continue d'appartie-
ment est spécifiée pour la procédure d'imputation, par exemple dans le cas où un(e) répon-
dant(e) aurait déclaré la valeur totale de ses tracteurs, mais non le nombre correspondant de
tracteurs. Tout le fichier des enregistrements donneurs acceptables est exploré et l'enregistre-
ment qui indique la plus petite différence entre les valeurs de la variable d'appariement est choisi
comme l'enregistrement donneur. Si deux ou plus de deux enregistrements donneurs éventuels
sont à la même distance, celui qui est géographiquement le plus près de l'enregistrement rece-
veur (selon les renseignements d'identification d'ordre géographique) est automatiquement choisi
comme enregistrement donneur. Si la population des enregistrements donneurs est vide, l'enre-
gistrement receveur est alors marqué d'un signe indiquant que l'imputation n'a pas réussi.

4.2.3 Type 3 — Recherches à plusieurs variables

Les recherches à plusieurs variables sont effectuées quand plus d'une variable continue d'appar-
tiement est spécifiée pour la procédure d'imputation. Il s'agit du type de recherche le plus com-
plexe parmi les trois types de recherche exécutés par le personnel chargé du traitement des don-
nées du recensement de l'agriculture de 1981. La méthode qu'utilise Statistique Canada pour
exécuter ce type de recherche a été mise au point par G. Sande.

Quand les données manquantes se rapportent à plus d'une variable continue d'appariement,
il vaut mieux utiliser comme enregistrement donneur l'enregistrement situé le plus près de l'enre-
gistrement receveur pour toutes les variables d'appariement en même temps. Cela requiert l'exé-
cution d'une recherche à plusieurs variables sur un grand fichier d'enregistrements donneurs
et a été rendu possible en pratique en regroupant la population des enregistrements donneurs
de telle sorte qu'il n'est pas nécessaire de scruter tous les enregistrements donneurs pour trou-
ver le plus proche. Un tel regroupement des enregistrements est appelé arbre de discrimination
(Key Discriminator (K-D) tree). On peut utiliser le même arbre "K-D" pour tous les enregistre-
ments nécessitant l'exécution d'une procédure donnée d'imputation "d'emprunt" dans une région
d'imputation donnée parce que le fichier des enregistrements donneurs acceptables est le même
dans chaque cas. Toutefois, si une procédure d'imputation "d'emprunt" différente doit être exé-
cutée sur une population d'enregistrements donneurs différente ou même si la même procédure
d'imputation "d'emprunt" doit être exécutée sur une région d'imputation différente, il faut cons-
truire un nouvel arbre "K-D" étant donné que le fichier des enregistrements donneurs accepta-
bles ne contiendra pas les mêmes enregistrements.

d'imputation "d'emprunt" ne peuvent être utilisés comme donneurs. Toutefois, les enregistrements dont une ou plusieurs valeurs doivent être imputées selon une procédure d'imputation interne ou d'emprunt. Les PI d'ordre 1 sont exécutées en premier, suivies des PI d'ordre 2 et 3 respectivement.

Pour faciliter le choix d'un enregistrement donneur acceptable, une ou plus d'une variable ne requérant aucune procédure d'imputation est choisie pour servir de variable d'appariement dans chaque procédure d'imputation "d'emprunt". Ces variables d'appariement, choisies par des experts en la matière, sont considérées comme fortement corrélées avec la (ou les) zone(s) dont une ou des valeurs doivent être imputées. L'enregistrement receveur et l'enregistrement donneur choisi doivent avoir des valeurs semblables pour les variables d'appariement choisies. Comme l'utilisation de variables d'appariement continues ne permet pas d'appariements exacts, on a recours à une fonction de distance reposant sur la (ou les) variable(s) d'appariement choisie(s) pour trouver l'enregistrement donneur acceptable le plus près du "mauvais" enregistrement. À chaque procédure d'imputation "d'emprunt" correspond un type de recherche parmi trois types possibles. La recherche par partition (type 1) est effectuée quand une seule variable discrète d'appariement est spécifiée pour la PI. La recherche dichotomique (type 2) est effectuée quand une seule variable continue d'appariement est spécifiée pour la PI. La recherche à plusieurs variables (type 3) est effectuée quand plus d'une variable continue d'appariement est spécifiée pour la PI. Chacun de ces trois types de recherche est décrit dans les paragraphes qui suivent. Il existe d'autres types de recherche faisant appel à d'autres combinaisons de variables d'appariement mais seuls les trois types mentionnés plus haut ont été utilisés pour le traitement des données du recensement de l'agriculture de 1981.

Enfin, une fois choisi l'enregistrement donneur acceptable, les valeurs des données de l'enregistrement donneur choisi dont on veut se servir pour faire l'imputation sont multipliées, si les renseignements de contrôle de la PI le prévoient, par un coefficient de proportionnalité avant d'être transférées à l'enregistrement receveur. Par exemple, si l'on utilise la variable "nombre de camions" comme variable d'appariement pour imputer la "valeur des camions", la valeur attribuée pour cette dernière à l'enregistrement receveur doit être égale à la "valeur des camions" de l'enregistrement donneur multipliée par le rapport du "nombre de camions" de l'enregistrement receveur sur le "nombre de camions" de l'enregistrement donneur. Comme on l'a mentionné précédemment, à chaque procédure d'imputation "d'emprunt" correspond un type de recherche parmi trois types possibles. Deux de ces types de recherche, les recherches dichotomiques et les recherches par partition, sont utilisés pour effectuer des procédures d'imputation pour lesquelles une seule variable d'appariement est spécifiée. L'autre type de recherche, les recherches à variables multiples, est utilisé quand plus d'une variable continue d'appariement est employée.

4.2.1 Type 1 — Recherches par partition

Les recherches par partition sont effectuées quand une seule variable discrète d'appariement pouvant prendre un petit nombre de valeurs est spécifiée pour la procédure d'imputation, par exemple dans le cas où un répondant aurait déclaré le nombre total de tracteurs, mais non la valeur correspondante en dollars. Comme il est peu probable qu'un exploitant agricole ait plus de 3 tracteurs, la population des enregistrements donneurs est divisée en parties: 1, 2 ou 3 tracteurs et plus. L'enregistrement donneur est choisi au hasard dans la partie des renseignements

une mesure fixée d'avance de la densité moyenne pour ce type d'arbre fruitier. On applique les procédures d'imputation interne en respectant certaines contraintes pour s'assurer que les valeurs imputées restent à l'intérieur de limites raisonnables.

L'application des procédures d'imputation interne est plus simple que celle des procédures d'imputation. Étant donné qu'une procédure d'imputation interne fait appel aux données du même enregistrement, on n'est pas obligé de spécifier un algorithme pour choisir l'enregistrement donneur comme on est obligé de le faire dans le cas des procédures d'imputation. Il suffit d'apporter la correction non aléatoire spécifiée par la procédure d'imputation interne appropriée à la zone (ou aux zones) dont on veut imputer la ou les valeurs. On exécute toutes les procédures d'imputation interne avant de procéder à l'imputation "d'emprunt".

4.2 Procédures d'imputation "d'emprunt"

Lorsque la ou les valeurs incohérentes d'une (ou de plusieurs) zone(s) ne dépendent pas de façon non aléatoire des valeurs cohérentes d'autres zones, les procédures d'imputation interne ne peuvent être appliquées. Le manque de renseignements en quantité suffisante sur l'enregistrement rejeté pour apporter une correction non aléatoire à la valeur ou aux valeurs incohérentes d'une zone (ou de plusieurs zones) oblige à recourir à une méthode d'imputation faisant appel aux données contenues dans un autre enregistrement. Cette méthode, appelée imputation "d'emprunt", suppose le transfert de données d'un "bon" enregistrement donneur (c'est-à-dire d'un enregistrement qui a passé toutes les vérifications obligatoires) à l'enregistrement rejeté. Les données transférées établiront la cohérence de la ou des données de la zone (ou des zones) de l'enregistrement rejeté dont la ou les valeurs sont incohérentes. Par exemple, on exécutera une procédure d'imputation "d'emprunt" pour estimer le nombre de têtes par type de bétail lorsqu'il n'y a que le nombre total de têtes de bétail qui est déclaré sur le questionnaire. Dans ce cas, la répartition par type de bétail qui est indiquée sur l'enregistrement donneur est imputée à l'enregistrement rejeté (receveur).

Étant donné que l'imputation "d'emprunt" requiert un algorithme pour trouver un enregistrement donneur, elle est plus complexe que l'imputation interne. Pour faire une imputation "d'emprunt", il faut spécifier un certain nombre de "paramètres" de recherche. Pour s'assurer que le "bon" enregistrement donneur est proche géographiquement du "mauvais" enregistrement receveur, le pays est divisé en régions géographiques distinctes appelées régions d'imputation. Le découpage des régions d'imputation correspond à la délimitation actuelle des "districts de culture" qui est définie en fonction de caractéristiques telles que le type de sol et le climat. Il y a, au Canada, 59 districts de culture et donc 59 régions d'imputation comprenant en moyenne 5 500 fermes. Pour être susceptible d'être choisi comme enregistrement donneur, un enregistrement doit obligatoirement provenir de la même région d'imputation que l'enregistrement receveur.

Pour éviter la recherche d'enregistrements qui ne peuvent pas fournir de données acceptables, chaque procédure d'imputation "d'emprunt" spécifie également la sous-population sur laquelle se fera la recherche de l'enregistrement donneur. Par exemple, si c'est la répartition du nombre de têtes par type de bétail qu'on veut imputer, les seuls enregistrements sur lesquels on ferait des recherches pour trouver un enregistrement donneur seraient les enregistrements faisant partie de la sous-population des enregistrements correspondant aux exploitants agricoles ayant déclaré posséder du bétail. Un enregistrement donné peut faire partie de plus d'une sous-population parmi les 30 qui existent. Dans certains cas, il se peut que tous les "bons" enregistrements de la région d'imputation soient considérés comme des donneurs acceptables; toute la population de la région d'imputation est alors définie comme la sous-population à considérer.

La dernière contrainte concernant le fichier des donneurs acceptables réside dans le fait que les enregistrements dont une ou plusieurs valeurs doivent être imputées selon une procédure

et marquées d'un signe indiquant qu'elles sont acceptables. Bien qu'aucune modification n'ait été apportée aux données de l'enregistrement du FMCI, cette opération est appelée 'ajuste-

ment de force' des données. Les rejets à une vérification obligatoire ne sont pas traités tout à fait comme les rejets à une vérification non obligatoire. Pour résoudre un rejet à une vérification obligatoire, on trans-met l'enregistrement rejeté à un commis responsable de la correction des rejets qui procède d'abord comme pour la résolution des rejets lors d'une vérification non obligatoire. Si toutefois il n'est possible de trouver aucune explication du rejet, au lieu d'"ajuster de force" les données de l'enregistrement rejeté, on marque l'enregistrement d'un signe indiquant qu'il doit être ache-miné pour traitement à la phase d'imputation informatique.

Comme dans le CC #1, il y a un cycle continu entre le personnel chargé de la correction des rejets et le programme de vérification du CC #2. Après chaque essai tenté pour résoudre un rejet lors d'une vérification du CC #2, l'enregistrement du FMCI est de nouveau soumis au traitement du programme de vérification du CC #2. Toutefois, contrairement au cas du CC #1, le commis à la correction des rejets n'a que trois essais pour résoudre les rejets lors d'une vérification du CC #2 d'un enregistrement donné du FMCI. Après le troisième essai, le pro-gramme de vérification du CC #2 est exécuté une dernière fois. Les rejets sur vérifications obligatoires qui persistent sont marqués du signe "ajustement de force" et les rejets sur vérifi-cations obligatoires qui persistent sont marqués du signe "imputation". Les rejets sur vérifica-tions obligatoires sont seulement indiqués à ce stade: les zones dont les valeurs doivent être imputées sont identifiées au cours de l'étape de l'imputation.

4. IMPUTATION

Le système d'imputation (voir figure 3) du recensement de l'agriculture de 1981 a pour objet de corriger les enregistrements de données du FMCI dont une ou plus d'une donnée aurait été rejetée à la vérification. Les rejets sur vérifications non obligatoires étant "ajustés de force", comme on l'a vu dans le chapitre qui précède, il ne reste que les rejets sur vérifications obliga-toires à résoudre à l'aide du système d'imputation. Pour que les enregistrements de données du FMCI ne soient pas rejetés sur vérifications obligatoires, des "procédures d'imputation" précises sont exécutées. Ces procédures d'imputation (PI), dont il y a plus d'une centaine, sont conçues de façon qu'il y ait le moins de zones possible dont les valeurs soient modifiées sur les enregistrements du FMCI; par exemple, les totaux sont toujours ajustés pour égaliser la somme des éléments qui les constituent, au lieu que ce soit les éléments qui sont ajustés pour égaliser leur somme. À chaque PI est associé un ensemble approprié de renseignements de contrôle et le choix d'une PI est déterminé par la ou les zones dont la ou les valeurs doivent être impu-tées. Il y a deux types de PI: les procédures d'imputation interne (ou corrections non aléatoires) et les procédures d'imputation dite "d'emprunt";

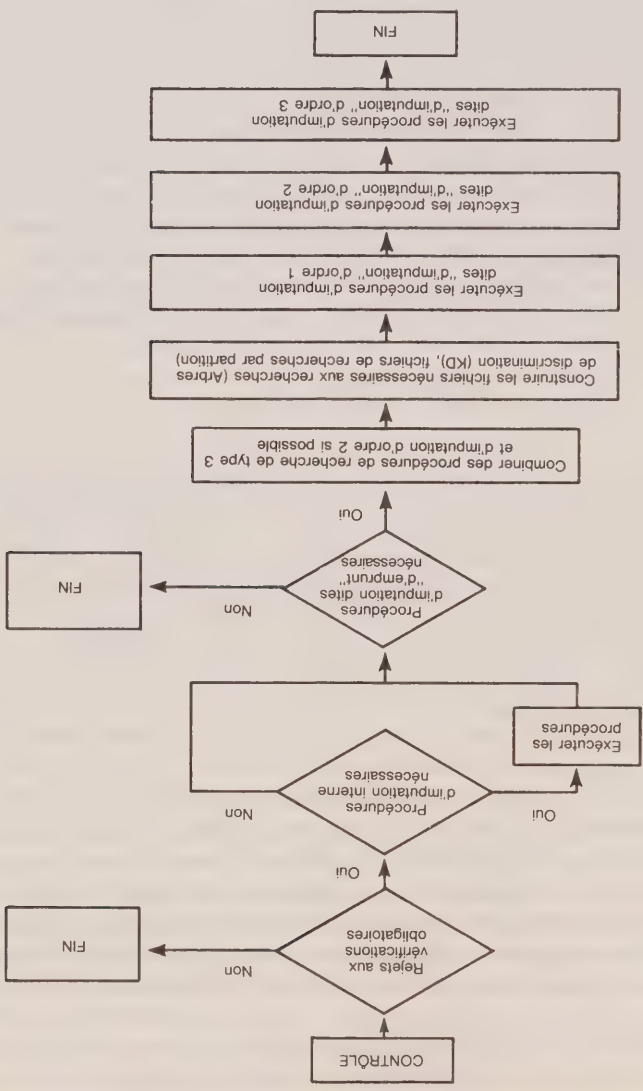
4.1 Procédures d'imputation interne

Les procédures d'imputation interne sont exécutées dans les cas où il existe suffisamment de données dans les enregistrements rejetés pour permettre au système d'imputation d'apporter une correction non aléatoire à la zone (ou aux zones) dont la valeur ou les valeurs sont incohé-rentes. Ces procédures d'imputation interne sont exécutées dans les cas où la ou les zones dont la ou les valeurs sont incohérentes dépend (ou dépendent) de façon non aléatoire d'autres zones dont les valeurs n'ont pas besoin d'être imputées. Par exemple, une procédure d'imputation interne serait exécutée dans le cas où un répondant aurait déclaré des quantités pour divers types de bétail, mais aurait omis de déclarer le nombre total de têtes. Dans ce cas, le nombre total de têtes serait la somme des quantités déclarées pour les divers types de bétail. Un autre cas dans lequel une procédure d'imputation interne serait exécutée est le cas où un répondant aurait déclaré une certaine quantité d'un type donné d'arbre fruitier, mais aurait omis d'indiquer la superficie correspondante. Dans ce cas, la superficie serait calculée en faisant intervenir dans les calculs

n'égale pas la somme des valeurs déclarées pour chacun des différents types de détail, il y aura rejet de l'enregistrement. Les vérifications obligatoires les plus complexes sont celles qui sont effectuées pour la section du questionnaire se rapportant aux cultures.

Pour résoudre un rejet à une vérification non obligatoire, l'enregistrement est transmis à un commis responsable de la correction des rejets. Celui-ci vérifie d'abord si le rejet à la vérification est attribuable ou non à une erreur de frappe. Si c'est le cas, la donnée pertinente est réintroduite. Dans le cas contraire, le commis scrute le questionnaire pour voir si le répondant n'a pas écrit sur le questionnaire qu'il a rempli des commentaires quelconques qui pourraient expliquer la cause du rejet. Par exemple, si le répondant est prié de répondre à une question en tonnes et si la réponse est exprimée en livres au lieu d'en tonnes, la réponse sera probablement rejetée lors d'une vérification non obligatoire de vraisemblance. Dans ce cas, le commis responsable de la correction des rejets convertira les livres en tonnes. Si le commis ne peut trouver aucune explication du rejet, les réponses sont laissées inchangées sur l'enregistrement du FMC

Figure 3. Schéma des opérations de l'étape d'imputation



CC #1 et CC #2 sont corrigés manuellement. Les rejets à la vérification qui n'ont pu être corrigés au cours du cycle de correction des rejets sont acheminés pour traitement à la phase d'imputation.

Tous les enregistrements du FMCI sont traités individuellement d'un bout à l'autre du système de contrôle.

3.1 Cycle de correction #1 (Vérifications à des fins de décodage et vérifications des renseignements d'identification)

Le cycle de correction #1 comprend l'application et la résolution de deux séries de vérifications: les vérifications à des fins de décodage et les vérifications des renseignements d'identification.

Les vérifications à des fins de décodage sont les premières auxquelles on procède et s'il existe des conditions qui empêchent le "désenchaînement" d'un enregistrement de données, il y aura rejet de l'enregistrement. Par exemple, comme deux zones ne peuvent avoir les mêmes caractères d'identification, le "désenchaînement" ne pourra se faire si l'on a introduit deux noms de zone identiques.

Tous les rejets à la vérification à des fins de décodage sont résolus manuellement par le personnel responsable de la correction des rejets. Cette opération suppose qu'on revienne au questionnaire pour déterminer la cause du rejet, puis qu'on réintroduise les données pertinentes. Après une première tentative pour résoudre un rejet à la vérification à des fins de décodage, l'enregistrement du FMCI est revérifié en étant soumis de nouveau à des vérifications à des fins de décodage, ces opérations formant un cycle continu entre les vérifications à des fins de décodage et le personnel responsable de la correction des rejets. Ce cycle est répété jusqu'à ce qu'il n'y ait plus aucune donnée de l'enregistrement du FMCI qui soit rejetée à la vérification à des fins de décodage. Si un rejet à la vérification à des fins de décodage ne peut être résolu directement, l'interprétation valide la plus appropriée des données disponibles est utilisée comme substitut final.

Une fois résolus tous les rejets à la vérification à des fins de décodage, on passe à l'étape des vérifications des renseignements d'identification. Si l'un ou l'autre renseignement d'identification figurant sur un enregistrement du FMCI est incohérent ou manquant, il y aura alors un ou plus d'un rejet à la vérification des renseignements d'identification. Ces rejets à la vérification des renseignements d'identification sont résolus de la même façon que le sont les rejets à la vérification à des fins de décodage.

Une fois que le personnel responsable de la correction des rejets a fini de résoudre tous les rejets à la vérification à des fins de décodage et à la vérification des renseignements d'identification du CC #1, l'enregistrement du FMCI est traité par le programme de vérification CC #2.

3.2 Cycle de correction CC #2 (Vérifications des données)

Les vérifications des données (CC #2) servent à détecter les erreurs qu'il pourrait y avoir dans le corps du questionnaire, par opposition aux erreurs de codage ou aux erreurs dans les renseignements d'identification. Il y a deux types de vérifications des données: les vérifications non obligatoires (75) et les vérifications obligatoires (24).

Les vérifications non obligatoires sont effectuées pour détecter les entrées douteuses dans les enregistrements de données du FMCI. Généralement, les vérifications non obligatoires, qui servent à détecter les valeurs de variables situées en dehors des limites prescrites, consistent à comparer des zones ou des groupes de zones d'un questionnaire donné pour déterminer si les valeurs de certaines données ne seraient pas anormalement élevées ou basses par rapport aux valeurs d'autres données. Par exemple, un enregistrement indiquant qu'une exploitation agricole a une superficie totale de 10 acres et compte 10 000 têtes de bétail signalé lors d'une vérification non obligatoire de vraisemblance.

Les vérifications obligatoires visent à détecter les impossibilités logiques qui pourraient exister dans les enregistrements de données; par exemple, si le nombre total de têtes de bétail déclaré

correction sur les enregistrements imparfaits ou, si la chose n'est pas possible, de faire passer les enregistrements encore imparfaits à la phase de l'imputation où les données imparfaites sont ajustées. On trouvera à la figure 2 un organigramme des opérations de la phase de contrôle.

Le système de contrôle est constitué de trois éléments: deux cycles de vérifications informatiques appelés cycles de correction #1 et #2 et un cycle de correction des rejets à la vérification, appelé correction des rejets. Le cycle de correction #1 (CC #1) comprend les vérifications qui détectent les conditions qui empêchent le "désenchaînement", c'est-à-dire qui empêchent la transformation d'un enregistrement d'un format enchaîné à un format fixe (ce sont les vérifications à des fins de décodage) et les vérifications qui détectent les erreurs au niveau des renseignements d'ordre géographique et des renseignements d'identification figurant sur la page couverture du questionnaire (ce sont les vérifications des renseignements d'identification). Le cycle de correction #2 (CC #2) comprend les vérifications qui déclèlent les incohérences au niveau des données du corps du questionnaire (ce sont les vérifications des données). La correction des rejets est une opération de commis au cours de laquelle les rejets à la vérification des cycles

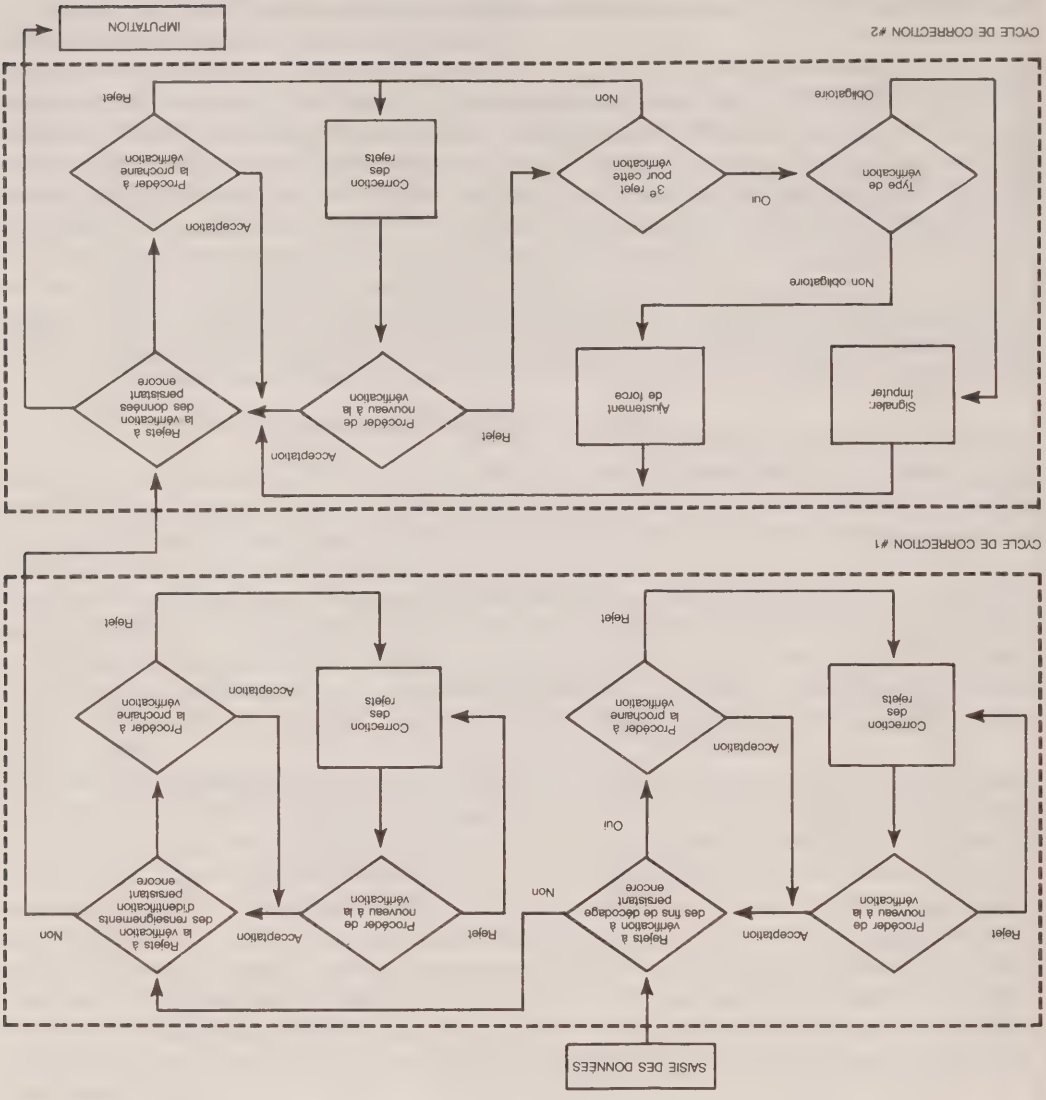


Figure 2. Organigramme des opérations de la phase de contrôle

aux règles établies pour les vérifications informatiques faites pendant la phase de contrôle. Les méthodes utilisées dans chacune de ces trois phases de traitement sont décrites plus en détail dans les sections qui suivent. La figure 1 présente un schéma résumant les diverses étapes du traitement des données du recensement de l'agriculture de 1981.

Pour le recensement de l'agriculture de 1981, tous les exploitants agricoles au Canada devaient remplir le même questionnaire. Ce questionnaire a 8 pages et 134 questions. Les questions se rapportent à tous les aspects de l'agriculture, par exemple les types de culture, les stocks de bétail, le matériel employé et l'utilisation des terres. Les exploitants agricoles devaient répondre seulement aux parties du questionnaire qui s'appliquaient à eux.

Comme on n'offre ici qu'une vue d'ensemble, il n'est pas possible de s'étendre sur les aspects techniques du traitement informatique des données du recensement de l'agriculture. Ces aspects sont traités en détail dans Shields et Vippong (1981), source sur laquelle est fondé notre exposé.

2. SAISIE DES DONNÉES

À l'étape de la saisie des données, les données du recensement de l'agriculture sont transférées des questionnaires originaux à un fichier sur un support informatique. La saisie des données comporte elle-même deux étapes: un pré-traitement manuel (examen initial) et l'entrée des données par terminal.

Une fois parvenus au bureau central pour y être traités, les questionnaires sont soumis à un pré-traitement dit examen initial. Dans ce processus, un commis scrute chaque questionnaire pour y déceler des irrégularités dans les réponses, par exemple des réponses illisibles, des réponses indiquant "idem" et des réponses inscrites au mauvais endroit. Si les réponses valides peuvent être discernées, elles sont enregistrées aux bons endroits; sinon, le questionnaire est laissé tel quel.

Ensuite, au cours de cette même phase de la saisie des données, les données de tous les questionnaires reçus sont introduites par clavier dans l'ordinateur. Les renseignements d'identification de la page couverture du questionnaire sont introduits selon un format standard fixe. Mais comme les exploitants agricoles doivent répondre seulement aux parties du questionnaire qui s'appliquent au type de leur exploitation agricole, une bonne partie du questionnaire demeure en blanc. Pour réduire le temps de saisie sur clavier, une méthode appelée "introduction en chaîne" est utilisée pour saisir le reste des données. D'après cette méthode, le nom de la zone est saisi par clavier, immédiatement suivi de la valeur de la donnée pour cette zone. Seules les zones pour lesquelles des valeurs de données existent sont saisies; les parties sans réponse du questionnaire ne le sont pas. En raison de l'éparpillement des réponses, cette méthode permet une économie appréciable du temps de saisie des données.

Le processus de saisie des données crée un enregistrement dans le fichier maître de contrôle et d'imputation (FMCI) pour chacun des questionnaires, qui sont au nombre d'environ 320 000. Il y a 244 zones sur un enregistrement du FMCI, chacune étant identifiée par un nom, en général de 6 caractères. L'opérateur responsable de l'introduction par clavier est chargé de taper "##" pour toutes les valeurs illisibles. Si la chose est possible, une correction est apportée par un commis sur les enregistrements contenant ce symbole au cours de la phase de contrôle; sinon, les enregistrements sont corrigés au cours de la phase d'imputation.

3. CONTRÔLE

La phase de contrôle vise deux objectifs. Le premier, c'est de procéder à des vérifications informatiques pour détecter toutes les entrées de données qui pourraient être incohérentes, man-

quantas ou suspectes. Le deuxième objectif, c'est de faire exécuter par des commis une

Méthode de traitement des données du recensement de l'agriculture de 1981

DAVID K. HOLLINS¹

RÉSUMÉ

Cet exposé présente une vue d'ensemble de la méthode utilisée pour le traitement des données du recensement de l'agriculture de 1981. L'accent est mis sur les méthodes de contrôle et d'imputation et, plus particulièrement, sur l'algorithme de recherche à plusieurs variables. Une brève évaluation de l'utilisation du système est également donnée.

MOTS CLÉS: Contrôle et imputation; recherches à plusieurs variables

1. INTRODUCTION

Cet exposé présente une vue d'ensemble de la méthode utilisée pour le traitement des données du recensement de l'agriculture de 1981. Le traitement des données comprend trois phases distinctes: la saisie des données, le contrôle et l'imputation, chacune a une fonction différente. Dans un premier temps, au moment de la saisie des données, les réponses aux questions du questionnaire du recensement sont introduites par terminal dans un fichier informatique. Ensuite, pendant la phase de contrôle, les enregistrements de ce fichier sont soumis à des vérifications informatiques visant à détecter toute entrée incohérente, manquante ou douteuse. Enfin, lors de l'imputation, certains enregistrements sont modifiés de manière à être rendus conformes

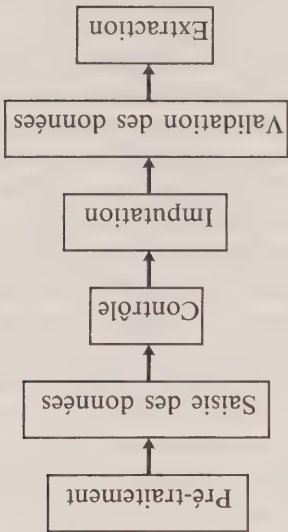


Figure 1. Schéma global des opérations de traitement

¹ D.K. Hollins, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, Parc Tunney
Ottawa (Ontario), Canada K1A 0T6.

Dans les domaines les plus petits, aucun des estimateurs non biaisés (EXT, POS, REG/T, REG/R) n'est avantageux du point de vue de la variance, surtout les estimateurs REG. Ce problème est résolu par les deux formes corrigées des estimateurs REG, les ERC.

C. Les deux formes de l'ERC, l'ERC/T et l'ERC/R, renferment un biais négligeable dans les cas où les estimateurs SYN sont presque sans biais (par exemple COMMERCE DE DÉTAIL, région 17); autrement, les ERC contiennent un certain biais qui, cependant, est généralement moins élevé que celui des estimateurs SYN (par exemple COMMERCE DE DÉTAIL, région 2). La variance et l'erreur quadratique moyenne des ERC sont beaucoup plus faibles, dans tous les domaines, que celles des estimateurs REG. Cette constatation est particulièrement fréquente dans les domaines les plus petits. En comparaison des autres estimateurs, on voit que les ERC (comme prévu) ont néanmoins une variance élevée dans pratiquement tous les domaines. Par contre, l'erreur quadratique moyenne des ERC est moins élevée que celle des estimateurs SYN dans les domaines où ces derniers sont extrêmement biaisés. Le tableau 6, par exemple, montre que l'erreur quadratique moyenne de l'ERC/R est plus faible que celle de l'estimateur SYN/R dans 9 petits domaines sur 16. Ce résultat découle évidemment du fait que, dans les domaines où l'estimateur SYN renferme un biais considérable, le carré du biais accroît énormément l'erreur quadratique moyenne de l'estimateur SYN, alors que le carré du biais n'est pas très important pour les ERC. Comme nous ignorons lesquels des domaines créent les biais élevés, il est généralement mieux d'utiliser les ERC pour obtenir des estimations fiables dans tous les domaines.

5. CONCLUSIONS

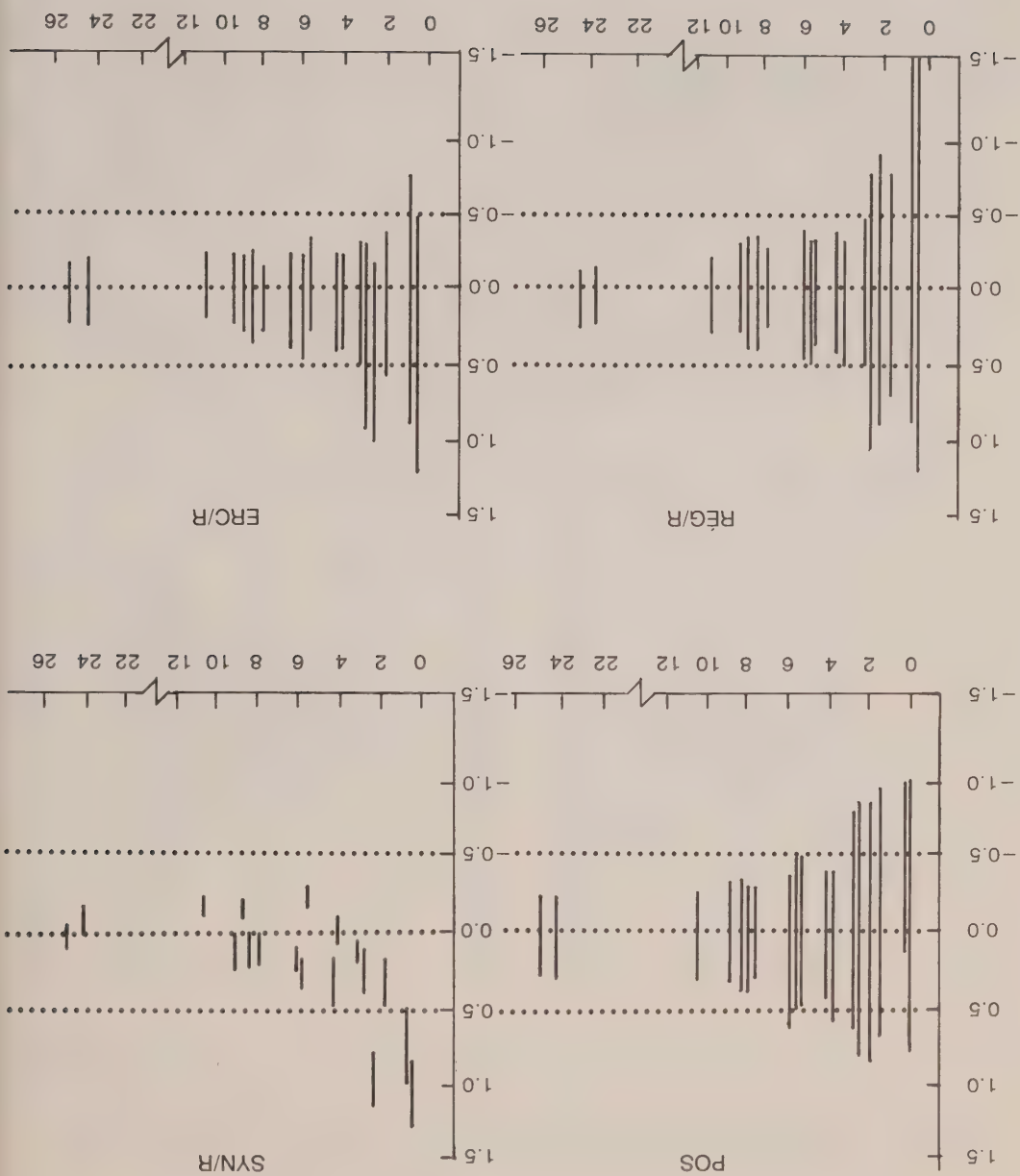
En somme, nous constatons que les ERC produisent des résultats qui permettent de les considérer comme des outils prometteurs pour les travaux futurs sur les estimations relatives aux petites régions. La méthode recommandée de calcul d'un intervalle de confiance à partir des ERC est décrite à la section 3.

Nous pensons que la méthode de calcul des ERC décrite plus haut représente un moyen simple de ramener les estimations un peu vers la valeur des estimateurs SYN, qui sont stables, quand le rendement de l'échantillon est moins élevé que prévu. Cet objectif (bien qu'atteint par des approches très différentes) est également la base de certains autres travaux récents, comme ceux fondés sur l'estimation empirique bayésienne (Fay et Herriot, 1979) et l'utilisation d'un estimateur dépendant de l'échantillon (Drew, Singh et Choudhry, 1982).

BIBLIOGRAPHIE

- DREW, J.D., SINGH, M.P., et CHOUDHRY, G.H. (1982). Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active du Canada. *Techniques d'enquête*, 8, 19-54.
- FAY, R.E. et HERRIOT, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- HIDIROGLOU, M.A., MORRY, M., DAGUM, E.B., RAO, J.N.K., et SÄRNDAAL, C.E. (1984). Evaluation of alternative small area estimators using administrative data. Communication présentée aux réunions de l'ASA, Philadelphie, août 1984.
- SÄRNDAAL, C.E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustments for nonresponse. *Bulletin of the International Statistical Institute*, 49:1, 494-513. (Actes de la 43^{ème} séance, Buenos Aires).
- SÄRNDAAL, C.E., et RÄBÄCK, G. (1983). Variance reduction and unbiasedness for small domain estimators. *Statistical Review*, 1983:5 (Essais en hommage à T.E. Dalenius), 33-40.
- SÄRNDAAL, C.E. (1984). Design-consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.

Figure 1 (continue)



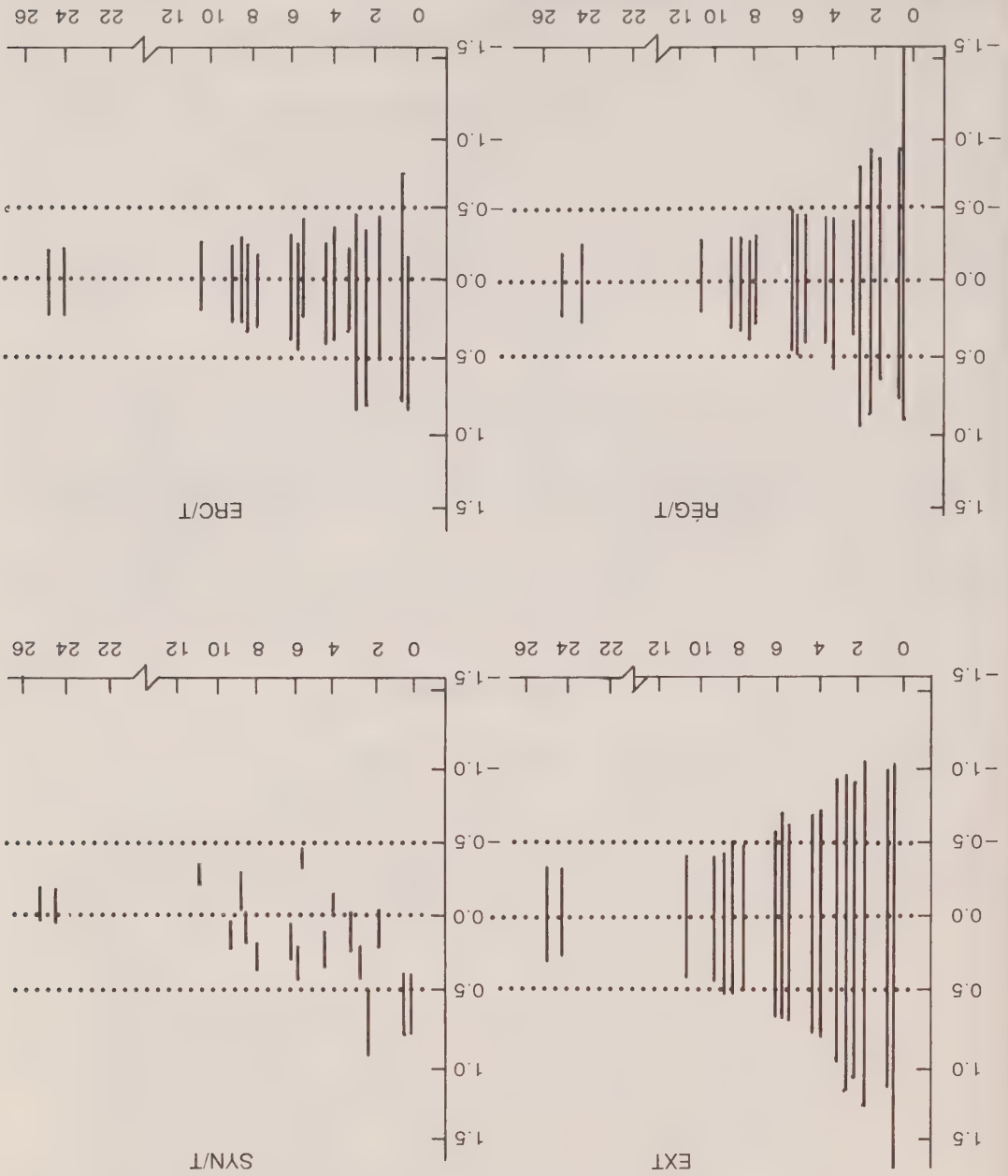


Figure 1: Secteur d'activité: COMMERCE DE DÉTAIL. Régions: 18 divisions de recensement en Nouvelle-Écosse. Bande de distribution de l'erreur relative des estimateurs - l'abscisse représente le rendement moyen de l'échantillon

Tableau 3
Rendement moyen de l'échantillon et biais relatif de chacun des huit estimateurs pour 500 échantillons aléatoires simples prélevés dans l'ensemble de la population
Secteur d'activité: HEBERGEMENT. Régions: 16 divisions de recensement en Nouvelle-Ecosse.

Région	Rendement moyen	EXT	POS	SYN/C	ERC/C	RÉG/C	SYN/R	ERC/R	RÉG/R
Estimateur									

1	0.25	0.01	-0.75	-0.08	-0.06	-0.01	0.36	0.28	0.01
2	1.37	-0.06	-0.21	0.25	0.10	0.02	0.25	0.11	0.02
3	1.02	0.06	-0.26	0.19	0.09	0.04	0.12	0.06	0.03
4	0.23	-0.10	-0.77	-0.33	-0.26	-0.07	-0.15	-0.13	-0.05
5	2.04	0.03	-0.13	0.21	0.08	0.03	0.18	0.06	0.01
6	1.49	0.04	-0.13	0.17	0.10	0.03	0.03	0.02	0.01
7	1.53	0.01	-0.18	-0.29	-0.11	-0.01	-0.30	-0.12	-0.02
8	1.54	0.03	-0.19	-0.42	-0.17	-0.01	-0.26	-0.11	-0.02
9	6.83	0.01	-0.02	0.13	0.02	0.00	0.12	0.02	0.00
10	1.26	-0.01	-0.26	0.40	0.17	0.03	0.30	0.13	0.02
11	3.06	0.04	-0.02	0.51	0.21	0.08	0.40	0.16	0.06
12	1.80	0.02	-0.16	-0.08	-0.05	-0.03	-0.23	-0.10	-0.03
14	1.04	0.02	-0.33	-0.52	-0.23	-0.07	-0.32	-0.15	-0.06
15	1.54	-0.03	-0.23	-0.21	-0.13	-0.08	-0.15	-0.11	-0.08
17	3.08	-0.07	-0.05	-0.03	-0.01	0.00	-0.14	-0.07	-0.03
18	0.52	0.04	-0.54	3.26	3.20	0.60	2.97	2.92	0.50

Tableau 4
Erreur quadratique moyenne de chacun des huit estimateurs pour 500 échantillons aléatoires simples prélevés dans l'ensemble de la population
Secteur d'activité: HEBERGEMENT. Régions: 16 divisions de recensement en Nouvelle-Ecosse.

Région	EXT	POS	SYN/C	ERC/C	RÉG/C	SYN/R	ERC/R	RÉG/R
Estimateur								
1	1,142	283	9	7	25	58	44	164
2	7,467	5,082	877	631	1,077	747	455	726
3	878	442	48	163	242	24	116	163
4	155	43	7	6	17	3	3	6
5	15,200	8,392	2,091	2,270	3,230	1,271	1,208	1,785
6	5,239	3,906	253	1,038	2,193	54	396	792
7	21,197	8,781	3,569	1,831	3,016	3,709	1,812	2,948
8	14,071	6,738	3,608	2,122	4,018	1,492	947	1,766
9	50,606	27,867	9,980	11,413	14,344	6,575	7,779	9,991
10	2,219	993	590	362	665	317	151	280
11	10,335	5,774	6,366	5,126	7,154	3,867	2,752	3,673
12	16,787	10,485	543	1,148	1,944	1,245	1,130	1,836
14	51,471	25,644	9,669	8,221	14,155	3,972	3,189	5,077
15	59,207	41,381	4,861	10,548	18,119	2,759	4,262	6,636
17	29,632	25,211	1,501	3,023	4,754	1,765	2,123	3,214
18	286	99	2,062	2,112	5,623	1,607	1,646	4,561

Tableau 1

Rendement moyen de l'échantillon et biais relatif de chacun des huit estimateurs pour 500 échantillons aléatoires simples prélevés dans l'ensemble de la population
Secteur d'activité: COMMERCE DE DÉTAIL. Régions: 18 divisions de recensement en Nouvelle-Ecosse.

Région	Rendement moyen	EXT	POS	SYN/C	ERC/C	RÉG/C	Estimateur	SYN/R	ERC/R	RÉG/R
--------	-----------------	-----	-----	-------	-------	-------	------------	-------	-------	-------

1	1.76	-0.02	-0.13	0.12	0.02	-0.03	0.30	0.09	-0.02	-0.02
2	5.45	0.00	-0.04	-0.36	-0.10	-0.02	-0.27	-0.08	-0.02	-0.02
3	3.90	-0.02	0.01	-0.08	-0.02	0.00	-0.01	-0.01	0.00	0.00
4	3.02	0.01	-0.05	0.15	0.05	0.01	0.13	0.04	0.04	0.04
5	5.93	0.00	0.01	0.21	0.05	0.00	0.13	0.03	0.00	0.00
6	7.63	-0.02	-0.01	0.28	0.07	0.01	0.10	0.02	0.00	0.00
7	8.61	0.02	0.01	-0.16	-0.03	0.01	-0.18	-0.03	0.01	0.01
8	5.64	-0.02	-0.01	0.34	0.10	0.03	0.24	0.06	0.01	0.01
9	24.64	0.00	0.00	-0.02	0.00	0.00	-0.01	0.00	0.00	0.01
10	8.92	-0.02	-0.02	0.15	0.02	-0.01	0.09	0.00	-0.01	-0.01
11	8.35	-0.03	-0.02	0.08	0.01	0.00	0.10	0.02	0.00	0.00
12	10.58	0.01	0.00	-0.27	-0.05	0.00	-0.18	-0.03	0.00	0.00
13	0.48	-0.04	-0.58	0.61	0.36	0.04	1.00	0.58	0.04	0.04
14	2.80	0.03	-0.03	0.33	0.11	0.00	0.24	0.10	0.02	0.02
15	4.21	0.06	-0.01	0.28	0.06	0.00	0.30	0.07	-0.01	-0.01
16	2.24	0.03	-0.05	0.74	0.26	0.03	0.94	0.32	0.02	0.02
17	23.95	-0.01	-0.01	-0.02	0.00	0.00	-0.05	-0.01	0.00	0.00
18	0.54	0.07	-0.54	0.63	0.34	-0.06	0.67	0.35	-0.06	-0.06

Tableau 2

Erreur quadratique moyenne de chacun des huit estimateurs pour 500 échantillons aléatoires simples prélevés dans l'ensemble de la population
Secteur d'activité: COMMERCE DE DÉTAIL. Régions: 18 divisions de recensement en Nouvelle-Ecosse.

Région	EXT	POS	SYN/C	ERC/C	RÉG/C	Estimateur	SYN/R	ERC/R	RÉG/R
--------	-----	-----	-------	-------	-------	------------	-------	-------	-------

1	3,209	2,206	96	697	1,397	462	769	1,484	1,484
2	42,598	24,623	21,782	12,725	17,338	13,110	10,256	14,380	14,380
3	10,469	6,853	357	2,592	4,212	146	2,333	3,782	3,782
4	5,626	3,657	324	746	1,186	257	1,206	1,853	1,853
5	14,554	9,681	2,999	5,090	7,360	1,294	3,993	5,974	5,974
6	12,308	5,686	6,713	3,423	4,289	1,255	1,747	2,515	2,515
7	34,865	17,988	6,912	9,387	13,451	8,161	12,019	17,239	17,239
8	12,066	8,630	5,772	3,694	5,045	2,981	3,528	4,986	4,986
9	72,974	40,440	5,776	24,025	29,250	5,068	21,292	25,832	25,832
10	22,091	9,433	4,559	5,832	7,927	2,009	5,365	7,272	7,272
11	23,519	12,505	1,778	6,738	9,578	2,348	7,890	11,063	11,063
12	46,588	21,874	35,310	13,558	17,084	17,454	12,222	16,514	16,514
13	635	244	161	95	228	422	287	783	783
14	3,871	2,849	692	1,254	2,141	378	1,373	2,346	2,346
15	8,088	3,511	2,249	1,892	2,806	2,651	1,985	2,937	2,937
16	3,245	2,127	3,316	1,563	2,516	5,333	1,741	2,654	2,654
17	81,211	47,753	5,503	28,957	35,232	7,681	27,457	33,136	33,136
18	1,003	306	169	187	654	186	184	637	637

4. RÉSULTATS DE L'ÉTUDE EMPIRIQUE

Pour étudier les propriétés des estimateurs décrits dans les sections précédentes, nous avons entrepris une simulation. La province de la Nouvelle-Ecosse a été choisie comme univers et la population comprend $N = 1,678$ unités (déclarations fiscales d'entreprises non constituées en société). La variable qu'on veut analyser est y , les salaires et traitements. Nous utilisons une seule variable auxiliaire, x , le revenu d'entreprise brut. On suppose que les valeurs de x_1, \dots, x_N sont connues.

La population a été classée en domaines par rapport à 4 secteurs d'activité économique et à 18 régions. Les secteurs d'activité économique sont: commerce de détail (515 unités), construction (496 unités), hébergement (114 unités) et autres activités économiques (553 unités). Les valeurs des coefficients de corrélation globaux entre les salaires et traitements et le revenu d'entreprise brut étaient 0.42 pour le commerce de détail, 0.64 pour la construction, 0.78 pour l'hébergement et 0.61 pour les autres activités économiques. Les régions sont les 18 divisions de recensement de la Nouvelle-Ecosse. On a obtenu 70 domaines non vides (il y avait 4 fois 18 domaines possibles, mais 2 ne comprenaient aucune unité). Ainsi, il faut estimer un total, t_d , pour chacun des 70 domaines chaque fois qu'un échantillon est prélevé. Pour la simulation de Monte Carlo, 500 échantillons aléatoires simples, s , de $n = 419$ unités chacun ont été tirés de la population de $N = 1,678$ unités. Les unités choisies pour l'échantillon ont été classées selon le secteur d'activité économique et la division de recensement. La population aurait également pu être divisée en fonction d'une deuxième dimension, par exemple selon la tranche de revenu. Mais pour les besoins de cette étude, on a supposé que tous les déclarants fiscaux sont compris dans une seule tranche de revenu ($G = 1$). Les résultats sont résumés pour chaque petite région dans les secteurs d'activité économique que COMMERCE DE DÉTAIL et HÉBERGEMENT sous forme de tableaux et de graphiques. Les tableaux 1 à 4 présentent le biais conditionnel relatif et l'erreur quadratique moyenne. La figure 1 est composée de 8 graphiques, dont un pour chacun des 8 estimateurs étudiés. Chaque graphique contient une "bande de distribution" verticale pour chacune des 18 divisions de recensement dans le secteur d'activité COMMERCE DE DÉTAIL. Les points maximum et minimum de chaque bande de distribution correspondent respectivement au 90^{ème} et au 10^{ème} centile de la distribution des 500 valeurs de $(t_d - t_d^s)/t_d^s$. Par conséquent, quand une bande de distribution est à peu près centrée sur le niveau horizontal 0, il s'ensuit que l'estimateur correspondant est presque sans biais pour le domaine en question, autrement l'estimateur est biaisé pour ce domaine. Plus une bande est courte, plus la variance de l'estimateur est faible dans un domaine.

Les tableaux et les graphiques permettent de tirer les conclusions suivantes (la conclusion C résume les résultats de cette étude et les conclusions A et B reposent sur les travaux de Särndal et Råbäck (1983) et Hidiroglou et coll. (1984)):

- A. Les estimateurs SYN/C et SYN/R sont extrêmement biaisés dans certains domaines, notamment ceux où le modèle implicite ne s'ajuste pas bien. Toutefois, un de leurs avantages est le fait que leur variance est toujours faible par rapport à celle des autres estimateurs. L'erreur quadratique moyenne des deux estimateurs SYN est donc très élevée dans les domaines où le biais est important (où la qualité de l'ajustement du modèle est mauvaise); par contre, leur erreur quadratique moyenne est faible dans les domaines où il y a peu de biais (où le modèle s'ajuste bien aux données).
- B. Les estimateurs REG/C et REG/R sont essentiellement non biaisés. Leur variance est généralement moins élevée que celle des estimateurs EXT et POS, mais elle est toujours beaucoup plus élevée que celle des estimateurs SYN/C et SYN/R.

L'ERC/C comportera un certain biais qui, cependant, est généralement beaucoup moins élevé que celui de l'estimateur SYN/C.
Les hypothèses sur lesquelles l'estimateur corrigé pour le calcul de ratios (ERC/R) est fondé sont, pour $g = 1, \dots, G$,

$$E^{\varepsilon}(Y^k) = \beta_g x_k; V^{\varepsilon}(Y^k) = \sigma_g^2 x_k, \quad k \in U_{.g}.$$

L'ERC/R est donc, dans le cas de l'échantillonnage aléatoire simple,

$$\hat{t}_{ERC/R}^g = \sum_{g=1}^G \{ X^{dg} \bar{R}_g + F_d N^{dg} (y^{sdg} - \bar{R}_g x^{sdg}) \} \quad (3.7)$$

où

$$\bar{R}_g = \frac{\sum_{d=1}^D N^{dg} \bar{y}^{sdg}}{\sum_{d=1}^D N^{dg} \bar{x}^{sdg}},$$

et

$$X^{dg} = \sum \bar{x}_k.$$

Drew, Singh et Choudhry (1982) ont élaboré des estimateurs pour petits domaines qui ressemblent aux estimateurs décrits dans cette étude. Leur estimateur à base de comptes est

$$\hat{t}_{dKNO/C}^g = \sum N^{dg} \{ W^{dg} \bar{y}^{sdg} + (1 - W^{dg}) \bar{y}^{s.g} \} \quad (3.8)$$

et leur estimateur à base de rapports est

$$\hat{t}_{dKNO/R}^g = \sum X^{dg} \left\{ W^{dg} \bar{y}^{sdg} + (1 - W^{dg}) \frac{\bar{x}^{s.g}}{\bar{y}^{s.g}} \right\} \quad (3.9)$$

où

$$W^{dg} = \begin{cases} \frac{n^{dg}}{E^{dg}} & \text{si } n^{dg} \leq E^{dg} \\ 1 & \text{autrement} \end{cases}$$

et $E^{dg} = n(N^{dg}/N)$. Dans le contexte de la présente étude, si on remplace W^{dg} dans les équations (3.8) et (3.9) par

$$W''^{dg} = \begin{cases} \left(\frac{E_d}{n_d} \right) \left(\frac{E^{dg}}{n^{dg}} \right) & \text{si } n_d < E_d \\ \left(\frac{E_d}{n_d} \right) \left(\frac{E^{dg}}{n^{dg}} \right) & \text{si } n_d \geq E_d \end{cases}$$

on obtient $\hat{t}_{ERC/C}$.

l'équation (3.1), et la probabilité d'une valeur négative est particulièrement élevée quand $n_d < E_d$. Avec le nouvel ERC, la possibilité d'obtenir des valeurs négatives est presque nulle. En pratique, si par un hasard peu probable une valeur négative est calculée pour l'ERC, nous recommandons que l'ERC soit redéfini comme étant égal à l'estimateur SYN qui est toujours positif.

Une formule naturelle pour estimer la variance de l'estimateur (3.2) est

$$V_p(t^{dsUB}) = \left(\frac{N_d}{N_d} \right)^2 \sum_{k \neq \ell}^{e_{sd}} \sum_{\ell} \Delta_{k\ell} \frac{\pi_k \pi_\ell}{(e_k - e_{sd})(e_\ell - e_{sd})} \tag{3.4}$$

où

$$e_{sd} = \frac{\sum_{sd} \frac{1}{\pi_k}}{\sum_{sd} \pi_k}$$

et

$$\Delta_{k\ell} = \begin{cases} 1 - \pi_k & \text{si } \ell = k \\ 1 - \frac{\pi_{k\ell}}{\pi_k \pi_\ell} & \text{si } \ell \neq k. \end{cases}$$

Nous soutenons que cette même formule peut bien servir à estimer la variance de l'ERC (3.3). Il est vrai que l'estimateur (3.3) est différent de l'estimateur (3.2) quand le rendement de l'échantillon est inférieur au rendement prévu; toutefois, il est peu probable que la différence entre ces deux estimateurs puisse être assez grande pour miner gravement la validité d'un intervalle de confiance de t_d centré sur t^{dERC} à partir de la variance estimée par l'équation (3.4).

Pour un échantillon aléatoire simple dans lequel, pour $g = 1, \dots, G$,

$$E_i(v_k) = \beta_g; V(v_k) = \sigma_g^2; k \in U_g, \tag{3.5}$$

on obtient la formule

$$\hat{\beta}_g = \frac{n_g}{\sum_{s,g} y_k} = y_{s,g},$$

qui permet de définir l'estimateur corrigé à base de comptes (ERC/C)

$$t^{dERC/C} = \sum_{g=1}^G \{ N^{dg}_{s,g} + F_d N^{dg}_{sdg} (y_{s,g} - y_{s,g}) \} \tag{3.6}$$

où E_d dans la formule pour F_d est maintenant

$$E_d = E_{vas}(n_d) = \frac{N}{n N_d}$$

et

$$N^{dg} \left(\frac{n}{N} \right)$$

Si le rendement prévu de l'échantillon dans le domaine d , $E_d = E^p(n_d) = \sum U_d \pi_k$, est élevé (par exemple si $E_d \geq 50$), il est alors pratiquement certain que le rendement réalisé, n_d , ne sera pas excessivement faible. Par exemple, dans l'cas, il est très rare d'obtenir des valeurs de $n_d \leq 30$. Dans ces cas, on peut recommander l'estimateur presque non biaisé (équation 3.2) sans aucune modification. Il devrait être beaucoup plus efficace que l'estimateur (3.1), notamment dans les domaines où l'ajustement du modèle n'est pas très bon. Toutefois, en pratique, il arrive souvent que certains domaines sont si petits que le rendement prévu de l'échantillon, E_d , ne dépasse pas 5. C'est ce qui s'est produit dans quelques-uns des domaines définis pour notre étude empirique. Dans ces cas, il est très probable que le rendement réalisé, n_d , oscille entre 0 et 5. Notre travail empirique a confirmé, ce qui est intuitivement évident, le fait que la correction des résidus pour ces petits domaines accroît beaucoup la variance, que cette correction soit exprimée sous forme directe, $\sum s_d e_k / \pi_k$, comme dans l'équation (3.1), ou sous forme de quotient, $N_d (\sum s_d e_k / \pi_k) / (\sum s_d 1 / \pi_k)$, comme dans l'équation (3.2).

Pour réduire cet accroissement de la variance, nous modifions le terme correctif de l'équation (3.2) d'une manière qui équivaut à accepter un biais faible (dans les domaines où l'ajustement du modèle n'est pas très bon) pour baisser l'accroissement de la variance quand le rendement de l'échantillon, n_d , est inférieur au rendement prévu (et on suppose également que le rendement prévu de l'échantillon est déjà faible à l'origine).

La forme du nouveau terme correctif est liée au rapport entre le rendement de l'échantillon, n_d , et le rendement prévu, E_d . Le terme correctif $\sum s_d e_k / \pi_k$ sera multiplié par (N_d / N_d) quand $n_d < E_d$ et par (N_d / N_d) autrement. Le terme correctif qu'on obtient avec cet "amortisseur" adaptable aura pour effet de ne pas "surcorriger" le terme synthétique si quelques-unes des erreurs résiduelles, e_k , ont des valeurs extrêmes quand n_d est faible. À cause de cette "surcorrection", on peut parfois sous-estimer de beaucoup le total pour un domaine d , obtenir des valeurs négatives lorsque seules des valeurs positives sont acceptables ou, à l'inverse, surestimer de beaucoup le total pour un domaine.

On obtient donc l'estimateur de régression corrigé (ERC), dont la définition varie selon que n_d est inférieur ou non à E_d :

$$(3.3) \quad t_{\text{ERC}}^d = \sum U_d y_k + F_d \sum \frac{e_k}{\pi_k} \quad \text{où} \quad F_d = \begin{cases} \frac{N_d}{N_d} & \text{quand } n_d \geq E_d \\ \frac{N_d}{N_d} & \text{quand } n_d < E_d \end{cases}$$

On peut démontrer que l'estimateur (3.3) est conditionnellement presque sans biais par rapport à n_d , pourvu que $n_d \geq E_d$. Si $n_d \leq E_d$, l'ERC renferme un biais conditionnel qui tend à augmenter plus n_d est au-dessous de sa valeur prévue. En même temps, l'ERC est poussé vers la valeur du terme synthétique, ce qui offre une certaine stabilité (variance faible). Globalement, l'ERC défini par l'équation (3.3) renferme un léger biais, mais à une variance beaucoup moins élevée que celle de l'estimateur REG.

Nous soulignerons un dernier avantage de l'ERC. À cause de sa variance considérable dans les domaines très petits, l'estimateur REG aura, avec une probabilité faible mais positive, des valeurs qui se situent extrêmement loin de la vraie valeur t_d . La valeur de l'estimateur REG peut même être négative, ce qui, naturellement, est inacceptable pour une variable (telle que les salaires et traitements) qui est par définition non négative. L'estimateur REG peut être négatif quand il existe d'importants résidus négatifs, e_k , dans le terme correctif de

3. ESTIMATEURS DE RÉGRESSION CORRIGÉS

Pour obtenir les estimateurs de régression présentés par Särndal (1984), on ajuste un modèle de régression à des variables auxiliaires et on calcule les valeurs prévues par le modèle pour les unités de chaque domaine de la population. Pour un plan de sondage arbitraire P (qui n'est pas nécessairement fondé sur l'éas) comportant des probabilités de sélection π_k (ordre premier) et π_{kt} (ordre deuxième), soit le modèle de régression

$$E_i(y_k) = \tilde{x}_k\tilde{\beta}; V_i(y_k) = v_k$$

où les y_k sont des variables aléatoires indépendantes. Un estimateur de $\tilde{\beta}$ est

$$\tilde{\hat{\beta}} = \left(\sum^s \tilde{x}_k\tilde{x}_k' \right)^{-1} \sum^s \tilde{x}_k'y_k \frac{v_k\pi_k}{v_k\pi_k}$$

où on suppose que les valeurs des v_k sont connues à un coefficient constant près et qu'on peut éliminer ce coefficient quand on calcule $\tilde{\hat{\beta}}$. Selon la méthode utilisée par Särndal (1984), la formule suivante produit un estimateur presque sans biais pour un total inconnu dans le domaine d :

$$(3.1) \quad t_{dREG}^{reg} = \sum^{U_d} y_k + \sum^{s_d} \frac{e_k}{\pi_k}$$

où $y_k = \tilde{x}_k'\tilde{\hat{\beta}}$ est la k^{leme} valeur prévue et $e_k = y_k - \tilde{y}_k$ représente la k^{leme} erreur résiduelle. Nous appellerons $\sum^{U_d} \tilde{y}_k$ le terme synthétique de l'estimateur t_{dREG}^{reg} et le deuxième terme, $\sum^{s_d} e_k/\pi_k$, le terme correctif. Si s_d n'est pas un ensemble vide, on peut substituer à l'estimateur REG (équation 3.1) un estimateur approximativement non biaisé:

$$(3.2) \quad t_{dSUB}^{SUB} = \sum^{U_d} y_k + N_d \frac{\sum^{s_d} \frac{e_k}{\pi_k}}{N_d}$$

où

$$N_d = \sum^s \frac{1}{\pi_k}$$

est l'estimation de la taille du domaine d . Le terme correctif peut maintenant être exprimé sous forme d'estimation par le quotient,

$$\frac{\sum^s \frac{e_k}{\pi_k}}{\sum^{s_d} \frac{1}{\pi_k}},$$

que l'on multiplie par la valeur connue de la taille du domaine d , N_d (bien entendu, N_d est connu puisque les totaux N_{d_g} sont connus). Comme la valeur de la taille n_d est aléatoire, l'estimation par le quotient permet de réduire la variance du terme correctif. Cette modification aura un effet particulièrement notable dans les domaines où la moyenne des résidus est nettement non nulle (c'est-à-dire dans les domaines où l'ajustement du modèle est médiocre).

où

$$y_{s.d.} = \sum_k \frac{y_k}{n_d}$$

est la moyenne des n_d valeurs de y dans le domaine d . Si $n_d = 0$, nous définissons l'estimateur POS comme nul (ce qui est assez arbitraire puisque, à proprement parler, cet estimateur est alors indéterminé). Ni l'estimateur EXT ni l'estimateur POS ne sont particulièrement avantagés. Ils servent surtout de repères pour comparer les propriétés d'autres estimateurs plus efficaces qu'on présente plus bas.

Deux formes des estimateurs SYN et REG ont été examinées, l'une pour le calcul de totaux, l'autre pour le calcul de ratios. L'estimateur SYN repose sur l'hypothèse selon laquelle un modèle donné s'applique à chaque groupe g . Pour le calcul de totaux, il est supposé que la moyenne de chaque groupe est identique dans tous les domaines d . Pour le calcul de ratios, il est supposé que le ratio entre une variable donnée et une variable auxiliaire est constant à l'intérieur d'un groupe particulier dans tous les domaines. Si cette hypothèse d'homogénéité des caractéristiques relatives aux domaines n'est pas exacte pour chaque groupe, les estimateurs SYN peuvent être extrêmement biaisés. L'estimateur REG proposé par Särndal (1984) permet a) de produire des estimations qui ne renferment presque aucun biais par rapport au plan de sondage et dont la variance est simple à estimer et les intervalles de confiance faciles à calculer (et raisonnables); b) de renforcer les estimations par l'inclusion de données de l'échantillon de tous les domaines.

Les formules pour le calcul de totaux à base de comptes sont:

Estimateur synthétique à base de comptes (SYN/C):

$$t_{\text{SYN/C}}^{d\text{SYN/C}} = \sum_{g=1}^G N_{dg} y_{s.g} \tag{2.4}$$

où $y_{s.g}$ est la moyenne de y dans $s.g$.

Estimateur de régression à base de comptes (REG/C):

$$\hat{t}_{\text{REG/C}}^{d\text{REG/C}} = \sum_{g=1}^G \{ N_{dg} y_{s.g} + N_{dg} (y_{s.dg} - y_{s.g}) \} \tag{2.5}$$

où $y_{s.dg}$ est la moyenne de y dans $s.dg$, et $N_{dg} = N n_{dg}/n$. Dans cette formule, $\sum_{g=1}^G N_{dg} (y_{s.dg} - y_{s.g})$ est un terme qui corrige le biais de l'estimation et entraîne en général un accroissement considérable de la variance.

Pour le calcul à base de rapports, les formules des estimateurs SYN et REG sont:

Estimateur synthétique à base de rapports (SYN/R):

$$\hat{t}_{\text{SYN/R}}^{d\text{SYN/R}} = \sum_{g=1}^G X_{dg} R_g \tag{2.6}$$

où $X_{dg} = \sum U_{dg} x_k$ et

$$R_g = \frac{\sum y_k}{\sum x_k}$$

Estimateur de régression à base de rapports (REG/R):

$$\hat{t}_{\text{REG/R}}^{d\text{REG/R}} = \sum_{g=1}^G \{ X_{dg} R_g + N_{dg} (y_{s.dg} - R_g x_{s.dg}) \} \tag{2.7}$$

moins élevée que celle de l'estimateur RÉG. En outre, l'erreur quadratique moyenne de l'ERC est inférieure à celle de l'estimateur SYN dans les domaines d'étude où ce dernier est extrêmement biaisé. Il est également facile d'établir des intervalles de confiance raisonnables pour le nouvel ERC.

Cet article se compose de cinq sections. La section 2 décrit quelques-uns des estimateurs qui sont utilisés le plus souvent pour le calcul d'estimations relatives à des petits domaines, notamment les estimateurs directs, les estimateurs pour les domaines stratifiés a posteriori et les estimateurs synthétiques, et quelques-uns des estimateurs de régression proposés par Särndal (1981, 1984). La section 3 présente des estimateurs de régression corrigés et en évalue les avantages et les inconvénients. Dans la section 4, on examine les propriétés des estimateurs de régression corrigés et de certains autres estimateurs à l'aide d'une simulation de Monte Carlo faite à partir de données fiscales des entreprises. Enfin, la section 5 présente quelques conclusions générales.

2. ESTIMATEURS

Soit une population $U = \{1, \dots, k, \dots, N\}$ divisée en D domaines d'étude distincts $U_1, \dots, U_d, \dots, U_D$. Soit N_d la taille de U_d . (Dans notre étude empirique, les domaines sont définis en fonction de 4 catégories d'activités économiques et des 18 divisions de recensement de la province de la Nouvelle-Ecosse. Il y avait $D = 70$ domaines non vides, comme l'ont décrit Hidiroglou, Morry, Dagum, Rao et Särndal (1984).)

La population se divise également selon une deuxième dimension en G groupes distincts, $U_1, \dots, U_g, \dots, U_G$. La taille de U_g est représentée par N_g . (Dans notre étude, les groupes correspondent à des tranches de revenu d'entreprise brut.) Si on classe les unités de la population selon le domaine et le groupe, on obtient DG catégories U_{dg} ; $d = 1, \dots, D$; $g = 1, \dots, G$. Soit N_{dg} la taille de U_{dg} .

La taille de la population, N , peut alors être exprimée de la manière suivante:

$$N = \sum_{d=1}^D N_d = \sum_{g=1}^G N_g = \sum_{d=1}^D \sum_{g=1}^G N_{dg} \tag{2.1}$$

Soit s un échantillon de taille n prélevé dans U par échantillonnage aléatoire simple (éas). Définissons s_d, s_g et s_{dg} comme les parties de s qui appartiennent, respectivement, à U_d, U_g et U_{dg} .

Les tailles correspondantes, qui constituent des variables aléatoires, sont représentées par n_d, n_g et n_{dg} . À noter que l'équation s'applique également aux n minuscules. La variable qu'on veut étudier, y (= salaires et traitements), a la valeur y_k pour la $k^{\text{ème}}$ unité (= déclaration fiscale d'une entreprise non constituée en société). La variable auxiliaire x (= revenu d'entreprise brut) vaut x_k pour la $k^{\text{ème}}$ unité et la valeur x_k est connue pour tous les $k = 1, \dots, N$.

On compare ici les estimateurs suivants du total pour l'ensemble d'un domaine, $t_d = \sum U_d y_k$ où $\sum U_d$ représente la sommation pour toutes les unités de U_d .

L'estimateur direct par extension (EXT) est:

$$t_{\text{EXT}} = \frac{n}{N} \sum s_d y_k \tag{2.2}$$

L'estimateur pour domaines stratifiés a posteriori (POS) est:

$$t_{\text{POS}} = N_d y_{s_d} \tag{2.3}$$

Étude empirique de quelques estimateurs de régression pour petits domaines

M.A. HIDIROGLOU et C.E. SÄRNDAI¹

RÉSUMÉ

La méthode classique d'estimation des caractéristiques d'un petit domaine d'étude est fondée sur l'utilisation de l'estimateur synthétique (SYN). L'avantage de cet estimateur est que sa variance est faible; par contre, il peut être très biaisé dans certains petits domaines dont la structure est différente de celle d'un ensemble de domaines. Särndal (1981) a proposé l'estimateur de régression (REG) pour le calcul d'estimations relatives à des domaines. Cet estimateur ne renferme presque aucun biais, mais il présente deux inconvénients: (i) sa variance peut être considérable dans certains petits domaines et (ii) il peut avoir des valeurs négatives dans certains cas où on ne peut pas en admettre.

Dans cette étude, nous proposons un estimateur intermédiaire qui représente un compromis entre les estimateurs SYN et REG. Cet estimateur, qu'on appelle "estimateur de régression corrigé" (ERC), a une variance beaucoup moins élevée que celle de l'estimateur REG et une erreur quadratique moyenne plus faible que l'estimateur SYN dans les domaines d'étude où ce dernier est extrêmement biaisé. L'ERC ne pose pas le problème des valeurs négatives mentionné plus haut. Ces propriétés sont vérifiées par une étude de Monte Carlo portant sur 500 échantillons.

MOTS CLÉS: Petits domaines; estimateur de régression; estimateur de régression corrigé; biais; erreur quadratique moyenne.

1. INTRODUCTION

L'avantage de l'estimateur synthétique (SYN) est sa faible variance; cependant, cet estimateur comporte les inconvénients suivants: a) il peut être extrêmement biaisé dans certains domaines d'étude et, en général, on ne peut déterminer les domaines dont il s'agit; b) par conséquent, un coefficient de variation (cv) ou un intervalle de confiance calculé pour ces domaines est inutile.

À l'aide du même modèle qui est à la base de l'estimateur SYN, on peut créer un estimateur analogue presque sans biais, l'estimateur de régression généralisé (REG), qui permet en outre de calculer, pour chaque estimation relative à un domaine, un intervalle de confiance selon la méthode classique fondée sur le plan de sondage. Un des inconvénients de l'estimateur REG, c'est que sa variance estimée (et donc son cv et l'étendue de son intervalle de confiance) peut être beaucoup trop élevée dans des domaines très petits (ce qui, bien entendu, découle directement du nombre insuffisant d'observations dans ces domaines). Par ailleurs, l'estimateur REG peut, bien que la probabilité de cette éventualité soit faible, avoir des valeurs négatives dans des cas où on ne peut pas en admettre.

Il paraît donc souhaitable de trouver le juste milieu entre l'estimateur SYN et l'estimateur REG. Nous présentons ici une étude empirique des propriétés d'un estimateur intermédiaire, l'estimateur de régression corrigé (ERC). Il renferme un biais faible (mais notable) dans les domaines où l'estimateur SYN est extrêmement biaisé; dans les autres domaines, l'ERC ne comporte presque aucun biais. Un autre avantage de l'ERC est que sa variance est beaucoup

¹ M.A. Hidiroglou, Division des méthodes d'enquête-entreprises, 5-C8, Immeuble Jean Talon, Parc Tunney, Université de l'Ontario) Canada K1A 0T6, et C.E. Särndal, Département de Mathématiques et Statistique, Université de Montréal, Montréal (Québec), Canada H3C 3J7.

- Quand la variable du khi-carré et l'erreur moyenne de prévision sont les seuls critères d'évaluation, le modèle 4 et 6 se classent aux premiers rangs.
- Quand on tient compte des huit critères d'évaluation, ce sont les modèles les plus simples (1 et 6) et le modèle 3 qui produisent les meilleurs résultats.
- Les modèles 1 (modèle à moyenne mobile) et 6 (modèle autorégressif) sont proches dans le classement non conditionnel (global), bien que le modèle 1 comprenne un paramètre de moins que le modèle 6.
- Dans le classement conditionnel, ces deux modèles se classent parmi les meilleurs, mais ils ne sont pas mutuellement exclusifs. Autrement dit, les modèles à moyenne mobile et les modèles autorégressifs sont complémentaires et nécessaires pour l'ajustement de séries chronologiques et le calcul de prévisions.
- Bien que le modèle 3 soit presque au dernier rang, il s'ajuste bien à une classe importante de séries chronologiques (séries qui renferment une tendance très forte) auxquelles tous les autres modèles s'ajustent mal.
- La somme des taux d'acceptation des modèles qui contiennent le plus grand nombre de paramètres (modèles 4 et 7) est de 61%, en comparaison d'un taux d'acceptation qui varie de 44% à 52% pour les modèles simples 1, 6 et 3.
- La somme des taux d'acceptation des modèles varie beaucoup d'un domaine à l'autre de l'économie. Elle s'élève à 93% pour les séries relatives au marché du travail, mais à seulement 21% dans le cas du commerce extérieur. Ce total dépend de la structure des séries chronologiques, des variations de cette structure et de l'importance de la composante irrégulière.
- Il semble que le classement conditionnel des modèles pour les extrapolations de valeurs à l'intérieur et à l'extérieur des échantillons dépende des phases du cycle économique ou de la conjoncture au moment où les séries prennent fin.

REMERCIEMENTS

Nous remercions M. Normand Lanier pour ses observations très pertinentes, Mme Helen Lim et M. Alfred Papineau pour leur précieuse collaboration technique et Mme B. Cohen pour le travail de dactylographie.

BIBLIOGRAPHIE

- BOX, G.E.P. et JENKINS, G.M. (1970). *Time Series Analysis Forecasting and Control*. San Francisco: Holden Day.
- BOX, G.E.P. et PIERCE, D.A. (1970). Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *Journal of the American Statistical Association* 65, 1509-1526.
- DAGUM E.B. (1980). *La méthode de désaisonnalisation X-11-ARMM*. n° 12-564F au catalogue, Statistique Canada, Ottawa.
- DRAPER, N.R. et SMITH, H. (1981). *Applied Regression Analysis*. John Wiley and Sons, Inc.
- HIGGINSON, J. (1976). A test for the presence of seasonality and a model test. Document de recherche, Division de la recherche et de l'analyse des chroniques, Statistique Canada, Ottawa.
- LJUNG, G.M. et BOX, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-307.
- PANDIT, S.M. et WU, S.M. (1983). *Time Series and System Analysis with Applications*. John Wiley and Sons, Inc.
- PLOSSER, C.I. et SCHWERT, G.W. (1977). Estimation of a non-invertible moving average process. *Journal of Econometrics*, 6, 199-224.
- PROTHERO, D.L. et WALLIS, K.F. (1976). Modelling macroeconomic time series (article suivi de commentaires). *Journal of the Royal Statistical Society*, AL39, 468-500.

Tableau 13 Pourcentages de rejets dus à l'erreur de prévision pour les extrapolations de valeurs à l'intérieur et à l'extérieur des échantillons

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	Modèle 7
(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)	(1, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 2)
A l'intérieur	34	35	41	32	34	34	33
A l'extérieur	31	32	42	33	31	32	31
	%	%	%	%	%	%	%

Tableau 14 Classements conditionnel et non conditionnel des modèles

Classement non conditionnel		Classement conditionnel	
Modèles	% de séries sans rejet	Modèles	% de séries sans rejet
1	40%	1	40%
6	28%	2	5%
5	27%	7	4%
2	20%	3	3%
3	14%		
7	10%		
4	2%		

Le tableau 14 est fondé sur les mêmes critères d'évaluation et les mêmes seuls critiques que ceux indiqués dans la deuxième colonne des tableaux 10 et 11. Le classement non conditionnel (global) de chaque modèle est exactement le même que dans la deuxième colonne du tableau 10. Seuls les taux d'acceptation des trois premiers modèles sont différents et, dans le tableau 14, le modèle 1 se révèle nettement supérieur aux autres modèles. Toutefois, le classement conditionnel de chaque modèle n'est pas le même que dans la deuxième colonne du tableau 11.

Les classements conditionnels présentés dans les tableaux 11 et 14 sont différents pour deux raisons. D'abord, bien entendu, le tableau 14 est fondé sur les extrapolations de valeurs à l'extérieur des échantillons. Une autre raison importante, c'est que le calcul des sept autres critères d'évaluation repose sur une année de données, et l'année manquante comprend une grave récession. Par conséquent, la structure des séries et l'évaluation des modèles produisent des résultats très différents.

Il semble donc que le classement conditionnel des modèles pour les extrapolations de valeurs à l'intérieur et à l'extérieur des échantillons dépende des phases du cycle économique ou de la conjoncture au moment où les séries prennent fin.

6. CONCLUSION

Notre objectif était de classer un ensemble de sept modèles ARMMI selon la qualité de l'ajustement et des prévisions obtenue pour un grand échantillon de séries chronologiques.

Le tableau 12 présente le classement conditionnel des modèles ARMMI pour les domaines de l'économie canadienne dans lesquels ces modèles ont été ajustés à douze séries chronologiques ou plus. Les critères d'évaluation et les seuils utilisés pour classer les modèles sont les mêmes que ceux indiqués dans la deuxième colonne des tableaux 10 et 11. On constate les choses suivantes:

- Les modèles 1 et 6 ont généralement les meilleurs résultats.
- La somme des taux d'acceptation des modèles varie beaucoup d'un domaine à l'autre; elle s'élève à 93% pour les séries relatives au marché du travail, mais à seulement 21% dans le cas du commerce extérieur.
- La somme des taux d'acceptation est d'au moins 50% dans cinq domaines. Ce total dépend de la structure des séries, des variations de cette structure et de l'importance de la composante irrégulière. Ce résultat est bon étant donné que, au cours de deux des trois dernières années des séries, le Canada a subi une sévère récession qui s'est répercutée fortement sur la structure des séries chronologiques. Le total des taux d'acceptation des séries relatives au commerce extérieur est toujours faible parce que ces séries sont très irrégulières.

5. EXTRAPOLATIONS DE VALEURS À L'INTÉRIEUR ET À L'EXTÉRIEUR DES ÉCHANTILLONS

Pour extrapoler des valeurs à l'intérieur d'un échantillon, on a ajusté les modèles à la totalité d'une série pour estimer les paramètres et calculer des extrapolations pour les trois dernières années. Les extrapolations de valeurs à l'extérieur d'un échantillon sont calculées sans tenir compte des renseignements d'après le début de la période d'extrapolation. Pour chaque point de départ de ce genre d'extrapolations, les paramètres ont été estimés de nouveau. Le tableau 13 montre les pourcentages de rejets dus à l'erreur de prévision au seuil critique de 15% pour les extrapolations de valeurs à l'intérieur et à l'extérieur des échantillons. L'écart entre ces deux pourcentages est faible et nettement inférieur à un écart-type pour chaque modèle. Le logiciel de désaisonnalisation X-11-ARMMI utilise des extrapolations de valeurs à l'intérieur d'un échantillon parce qu'elles coûtent moins que les extrapolations de valeurs à l'extérieur d'un échantillon.

Tableau 12
Classement conditionnel des modèles ARMMI dans différents domaines de l'économie canadienne

Secteurs	Rang des modèles et pourcentage des séries pour lesquelles le modèle est acceptable							
	1er rang	2e rang	3e rang	4e rang	1er rang	2e rang	3e rang	4e rang
Marché du travail	1	79	3	14	—	0	—	0
Prix	5	50	7	17	2	8	—	0
Industries manufacturières	3	19	6	14	1	5	2	5
Combustibles, énergie et exploitation minière	1	46	6	4	—	0	—	0
Commerce intérieur	1	53	6	7	7	7	—	0
Commerce extérieur	6	21	—	0	—	0	—	0
Transports	1	54	5	8	—	0	—	0
Finances	1	32	3	11	—	0	—	0

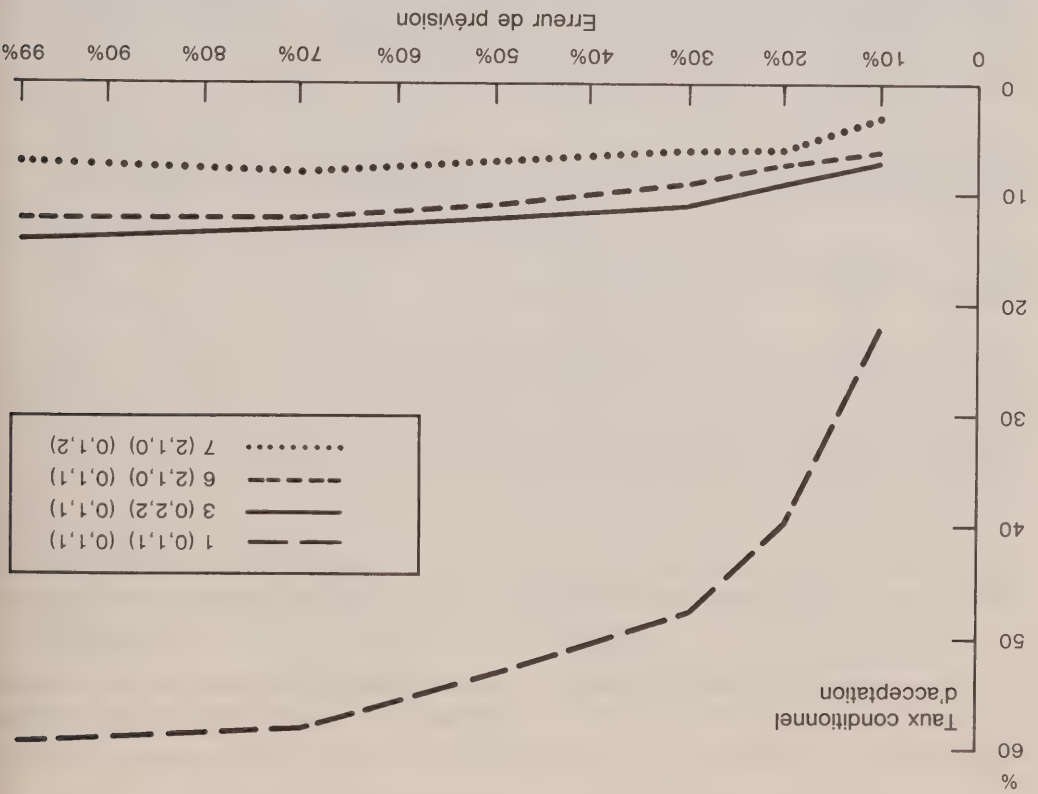


Figure 1. Classement conditionnel des modèles pour différents seuils critiques de l'erreur de prévision

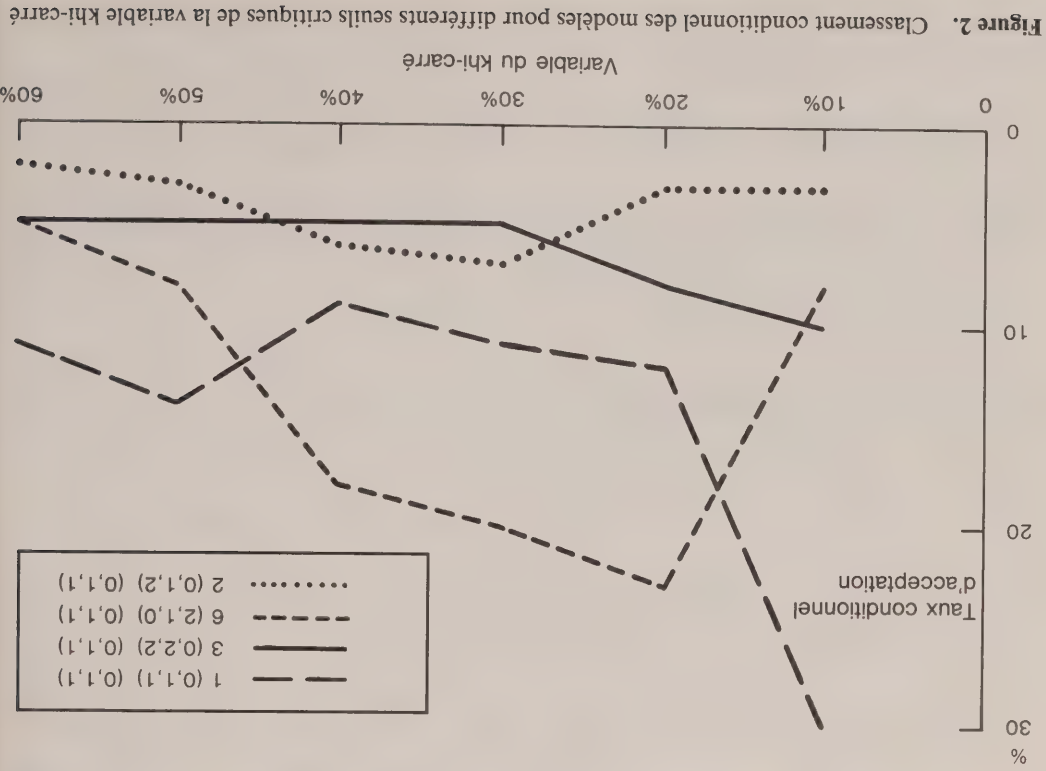


Figure 2. Classement conditionnel des modèles pour différents seuils critiques de la variable khi-caré

- Quand le modèle 6 ne se classe pas au premier rang, il suit de près au deuxième rang. Moins les seuils des critères sont stricts, plus le taux d'acceptation des modèles est élevé, quoique le classement des modèles demeure à peu près le même.

Le tableau 11 permet de constater les choses suivantes:

- Quand on tient compte de tous les critères d'évaluation, les modèles 1 et 6, qui occupent le premier et le deuxième rang dans le tableau 10, sont maintenant classés au premier et au troisième rang seulement.
- Le modèle 3 se classe au deuxième rang. Dans le tableau 10, ce modèle occupe tantôt le troisième, tantôt le cinquième et tantôt le sixième rang avec des taux globaux d'acceptation de 41%, 13%, 17% et 19% mais, dans le tableau 11, ce modèle se classe une fois au quatrième rang et trois fois au deuxième rang. Ce résultat est attribuable au fait que le modèle 3 s'ajuste bien à une classe importante de séries chronologiques (séries qui renferment une tendance très forte) auxquelles tous les autres modèles s'ajustent mal.

- Les modèles à moyenne mobile et les modèles autorégressifs ne sont pas mutuellement exclusifs. Ces deux genres de modèle sont complémentaires et nécessaires pour l'ajustement aux séries chronologiques et au calcul des prévisions.

- Quand les seuls critères d'évaluation sont que l'erreur moyenne de prévision soit inférieure à 15% et que la variable du khi-carré soit supérieure à 5%, la somme des taux d'acceptation des modèles 4, 7, 2 et 3 est de 63%.

- Quand on tient compte des huit critères d'évaluation, les modèles choisis sont simples et la somme de leur taux d'acceptation varie de 46% à 54% lorsque les seuils de l'erreur moyenne de prévision et de la variable du khi-carré sont respectivement de 15% et de 5%. Le taux d'acceptation dépend des seuils fixés pour les tests pour la présence de paramètres non significatifs et d'un trop grand nombre de différences.

- Le modèle 1 ne figure pas dans la troisième colonne du tableau 11, mais il y serait si le seuil établi pour l'erreur de prévision permise était porté à 20%.
- Les critères et les seuils de sélection des modèles indiqués dans les figures 1 et 2 sont les mêmes que ceux de la deuxième colonne des tableaux 10 et 11, sauf qu'à la figure 1 la valeur permise de l'erreur moyenne de prévision varie de 10% à 99% et qu'à la figure 2 le seuil de la variable du khi-carré varie de 10% à 60%.

La figure 1 révèle les choses suivantes:

- Ce sont les modèles 1, 3 et 6 qui ont les meilleurs résultats.
- Les rangs de classement des modèles demeurent généralement les mêmes.
- Le taux d'acceptation du premier modèle s'accroît plus rapidement que celui des autres modèles. Le taux d'acceptation du modèle 1 augmente de 23% à 59%, en comparaison de 13% à 17% pour le modèle 3. Il importe d'expliquer ce résultat. Le modèle 1 est choisi à cause de la valeur non conditionnelle (globale) de son taux d'acceptation, alors que les autres modèles sont choisis à cause de la valeur conditionnelle de ce taux.

- L'accroissement du taux d'acceptation des modèles dans le classement non conditionnel est plus prononcé que dans le classement conditionnel.

La figure 2 permet de tirer les conclusions suivantes:

- Les modèles 1, 3 et 6 sont généralement les meilleurs modèles pour n'importe quel seuil de la variable du khi-carré.

- L'ordre des modèles 1 et 6 est inversé à certains seuils critiques, mais ces modèles ne sont pas mutuellement exclusifs.

Tableau 10
Classement global des modèles

Modèles	Pourcentage	de séries	Modèles	Pourcentage	de séries	Modèles	Pourcentage	de séries	Modèles	Pourcentage	de séries
2 critères	$\chi^2 \geq 5\%$		8 critères*	$EP \leq 15\%$ $\chi^2 \geq 5\%$ $PF \leq 0.10$		8 critères*	$EP \leq 15\%$ $\chi^2 \geq 5\%$ $PF \leq 0.05$		8 critères*	$EP \leq 15\%$ $\chi^2 \geq 5\%$ $PF \leq 0.05$	
4	52%	1	34%	6	38%	6	38%	6	39%	1	39%
7	51%	6	31%	1	37%	1	37%	1	38%	2	38%
6	49%	5	23%	2	29%	2	29%	2	29%	5	29%
1	48%	2	20%	5	26%	7	26%	7	28%	3	27%
3	41%	7	11%	3	17%	3	17%	3	19%	4	19%
5	32%	4	2%	4	4%	4	4%	4	5%	5	5%

*Les valeurs des quatre critères autres que ceux indiqués ont été imposées.
EP = erreur de prévision; PF = paramètres non significatifs; TGND = trop grand nombre de différence

Tableau 11
Classement conditionnel des modes

Modèles	Pourcentage	de séries	Modèles	Pourcentage	de séries	Modèles	Pourcentage	de séries	Modèles	Pourcentage	de séries
2 critères	$\chi^2 \geq 5\%$		8 critères*	$EP \leq 15\%$ $\chi^2 \geq 5\%$ $PF \leq 0.10$		8 critères*	$EP \leq 15\%$ $\chi^2 \geq 5\%$ $PF \leq 0.05$		8 critères*	$EP \leq 15\%$ $\chi^2 \geq 5\%$ $PF \leq 0.05$	
4	52%	1	34%	6	38%	6	38%	6	39%	1	39%
7	9%	3	6%	3	9%	3	9%	3	9%	1	4%
2	1%	6	4%	7	4%	7	4%	7	4%	4	2%
3	1%	5	2%	2	3%	2	3%	2	2%	4	2%

*Les valeurs des quatre critères autres que ceux indiqués ont été imposées.
EP = erreur de prévision; PF = paramètres non significatifs; TGND = trop grand nombre de différences

Le tableau 10 révèle les choses suivantes:

- Quand la variable du khi-carré (χ^2) et l'erreur moyenne de prévision (EP) sont les seuls critères d'évaluation, les modèles 4 et 7, qui ont le plus grand nombre de paramètres, se classent aux premiers rangs.
- Par contre, quand on tient compte des huit critères d'évaluation, ce sont les modèles les plus simples (1 et 6) qui sont favorisés et ce à tous les seuils fixés pour les tests pour la présence de paramètres non significatifs (PF) et d'un trop grand nombre de différences (TGND).
- Les modèles 1 et 6 occupent généralement des rangs voisins, bien que le modèle 1 com-
porte un paramètre de moins que le modèle 6.

Tableau 8

SEUL		CRITIQUE	
CATÉGORIE I		CAT. II	
Modèle 1	(0, 1, 1) (0, 1, 1)	Modèle 4	(1, 1, 0) (0, 1, 1)
Modèle 2	(0, 1, 2) (0, 1, 1)	Modèle 5	(2, 1, 0) (0, 1, 1)
Modèle 3	(0, 2, 2) (0, 1, 1)	Modèle 6	(2, 1, 0) (0, 1, 1)
Modèle 7	(2, 1, 0) (0, 1, 2)	Modèle 7	(2, 1, 0) (0, 1, 2)
10	89	10	85
15	57	15	55
20	39	20	40
25	32	25	34
30	24	30	27
%	%	%	%

Tableau 9
Moyenne (M) et écart-type (ET) conditionnels de l'erreur moyenne de prévision

[illegible]

Le tableau 9 présente l'erreur moyenne de prévision et l'écart-type de cette erreur pour les modèles qui ont été acceptés et les modèles qui ont été rejetés à divers seuils de l'erreur de prévision. En plus d'avoir le taux de rejet le plus élevé, le modèle 3 affiche également l'erreur de prévision la plus importante quand il est rejeté. Les erreurs de prévision du modèle 3 sont aggravées par le fait que ce modèle renferme un trop grand nombre de différences. Toutefois, quand les erreurs de prévision du modèle 3 sont acceptables par rapport au seuil critique, leur moyenne est aussi faible que celle des erreurs des autres modèles.

4. CLASSEMENT DES MODÈLES SELON LEURS RÉSULTATS

Avant de classer les modèles, on a fixé différents seuils d'acceptation pour les huit critères d'évaluation. Les tableaux 10 et 11 présentent le classement global et conditionnel des modèles. Les pourcentages qui figurent au tableau 10 indiquent le taux global d'acceptation des modèles. La première ligne du tableau 11 présente le taux global d'acceptation du meilleur modèle; les autres modèles ont été choisis selon leur taux d'acceptation dans le cas des séries chronologiques pour lesquelles les modèles indiqués aux lignes précédentes ont été rejetés.

Tableau 6
Rejets dus à la présence de paramètres non significatifs

SEUIL CRITIQUE	CATÉGORIE I						CAT. II			CATÉGORIE III				
	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	Modèle 7	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	
.05	15	9%	49	29%	21	13%	42	25%	12	7%	22	13%	72	43%
.10	26	16%	88	53%	43	26%	73	44%	31	19%	45	28%	114	68%
	(0, 1, 1)	(0, 1, 1)	(0, 1, 2)	(0, 1, 1)	(0, 2, 2)	(0, 1, 1)	(2, 1, 2)	(0, 1, 1)	(1, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 1)	(2, 1, 0)	(0, 1, 2)

Tableau 7
Rejets dus à une corrélation entre les paramètres

SEUIL CRITIQUE	CATÉGORIE I			CAT. II			CATÉGORIE III		
	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	Modèle 7		
--	--	3	2%	86	51%	124	74%		
--	--	--	--	--	--	--	--		
(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)	(1, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 2)		

Une forte corrélation positive ou négative entre les estimations des paramètres n'est pas souhaitable et constitue un symptôme d'ambiguïté dans les valeurs estimées puisque les paramètres peuvent alors avoir différentes valeurs et produire des modèles dont la qualité de l'ajustement est égale. Le tableau 7 montre que seuls les modèles 2, 3 et 4 affichent des cas de rejet à cause du test de corrélation (c'est-à-dire que la valeur absolue d'au moins un des coefficients de corrélation était supérieure ou égale à 0,90). Ce problème est minime pour le modèle 2, mais grave pour les modèles 3 et 4, dont les paramètres ont manifesté une corrélation élevée pour 51% et 74% des séries auxquelles ces modèles ont été ajustés. Ce résultat est peut-être dû à un nombre trop élevé de différences dans le modèle 3 et à la présence d'un nombre excessif de paramètres dans le modèle 4.

3.2 Critère relatif aux extrapolations des modèles ARMMI

Ce critère a pour objet d'assurer la qualité des prévisions calculées à partir des modèles ARMMI. Nous voulons que le pourcentage moyen d'erreur de prévision soit au-dessous d'un niveau donné. Le tableau 8 révèle que six des sept modèles sont équivalents du point de vue de la qualité des prévisions; autrement dit, le nombre de paramètres autorégressifs et de paramètres de moyenne mobile n'a aucun effet sur l'erreur moyenne de prévision de ces modèles pour l'ensemble des séries chronologiques. Bien entendu, certains modèles produisent de meilleurs résultats pour certaines séries.

Tableau 4
Rejets dus à un nombre insuffisant de différences

SEUIL CRITIQUE	CATÉGORIE I			CAT. II			CATÉGORIE III		
	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	Modèle 7		
	(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)	(1, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 2)		
.90	--	--	--	14	8%	--	--	--	--

Dans cette étude, les seuils critiques établis pour décider si un modèle renferme un trop grand nombre de différences sont 0.90 et 0.95. Le tableau 5 montre que les modèles 3 et 4 affichent le plus grand nombre de cas de différences excessives. Le modèle 3 contient deux différences d'ordre un et deux paramètres de moyenne mobile dans sa partie non saisonnière. Si la deuxième différence d'ordre un n'est pas nécessaire, le modèle crée un certain degré d'autocorrélation si une série a déjà subi une différence d'ordre un. À cause de cette autocorrélation engendrée par le modèle, une des racines du polynôme de la moyenne mobile aura une valeur proche de un. On peut donc simplifier le modèle en éliminant un paramètre de moyenne mobile et une différence. Dans le cas du modèle 4, les rejets sont peut-être dus au nombre excessif de paramètres dans ce modèle.

Pour un modèle ARMMI d'un processus stochastique, il suffit d'examiner les deux premiers moments, soit la moyenne et le coefficient d'autocovariance. Le test de l'importance d'un paramètre sert uniquement à supprimer les paramètres dont l'apport à l'explication de l'autocovariance est faible ou nul.

Le tableau 6 montre deux choses. Premièrement les modèles les plus simples réussissent mieux à ce test que les modèles les plus complexes. Après une transformation logarithmique, la plupart des séries multiplicatives de l'échantillon évoluent à peu près en ligne droite (sauf pour les variations saisonnières), de sorte qu'un "modèle à différence d'ordre un" s'ajuste bien à ces données avec un petit nombre de paramètres. Si on ajoute au modèle un paramètre supplémentaire qui n'est pas vraiment nécessaire, la valeur estimée de ce paramètre sera sous-vent faible. Deuxièmement, les valeurs estimées des paramètres de moyenne mobile sont faibles (inférieures à 0.05 ou à 0.10) plus souvent que les valeurs estimées des paramètres autorégressifs. Par exemple, au seuil de 0.05, le deuxième paramètre autorégressif du modèle 6 est jugé inutile dans 13% des cas, en comparaison de 29% des cas pour le deuxième paramètre de moyenne mobile du modèle 2. De même, l'addition d'un deuxième paramètre de moyenne mobile dans la partie saisonnière a porté le taux de rejet de 13% pour le modèle 6 à 43% pour le modèle 7.

Tableau 5

Rejets dus à un trop grand nombre de différences

SEUIL CRITIQUE	CATÉGORIE I			CAT. II			CATÉGORIE III							
	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	Modèle 7							
	(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)	(1, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 2)							
.90	8	5%	11	7%	43	26%	50	30%	7	4%	9	5%	14	8%
.95	3	2%	6	4%	19	11%	37	22%	3	2%	3	2%	6	4%

Deux explications sont possibles. D'abord, ce modèle est parfois non inversible sans être rejeté pour les autres critères d'évaluation. Dans d'autres cas, ce modèle est à la fois non inversible et non stationnaire. Le fait que les racines de la partie non autorégressive peuvent être voisines de un peut causer une autocorrélation entre les résidus. Les valeurs des paramètres de la moyenne mobile deviennent alors plus élevées pour compenser cette autocorrélation. Un des critères efficaces pour évaluer un modèle ARMMI appliqué à une série chronologique- que est la variable du khi-carré de Box et Pierce (1970) (qui a été modifiée par Prothero et Wallis en 1976 et Ljung et Box en 1978), qui permet de vérifier s'il existe une autocorrélation entre les résidus. Le tableau 3 indique le nombre et le pourcentage de séries pour lesquelles chaque modèle a été rejeté à différents seuils critiques du test khi-carré. Ce tableau révèle deux choses: premièrement, dans une catégorie donnée de modèles, les modèles les plus simples ont les taux de rejet les plus élevés et, deuxièmement, le taux de rejet est dans une grande mesure lié à la catégorie à laquelle un modèle appartient. Pour illustrer la première constatation, nous notons que les modèles à moyenne mobile semblent être acceptés par le test khi-carré plus souvent que les modèles autorégressifs. Ce résultat est peut-être imputable à la présence de valeurs extrêmes dans les séries chronologiques. Au seuil de 5%, par exemple, le modèle I est rejeté pour 27% des séries, en comparaison de 49% pour son homologue autorégressif, le modèle 5. Les modèles de la catégorie III et celui de la catégorie II, le modèle mixte, sont tous inférieurs au deuxième modèle de la catégorie I.

Le nombre de différences dans un modèle est insuffisant lorsqu'une racine de l'équation caractéristique du polynôme d'autorégression est voisine de un, admettons par une marge de ξ . Dans cette étude, la valeur de ξ est fixée à 0.1. On peut voir au tableau 4 que seul le modèle 4 comporte des rejets dus à un nombre insuffisant de différences. Ce résultat peut être attribué au nombre excessif de paramètres dans ce modèle. Le modèle 4 contient deux paramètres autorégressifs et deux paramètres de moyenne mobile dans sa partie non saisonnière. Il existe une probabilité modérée qu'au moins une des estimations des paramètres autorégressifs puisse être égale ou supérieure à 0.9.

Tableau 3
Rejets dus au résultat du test khi-carré

SEUIL CRITIQUE	CATÉGORIE I										CAT. II										CATÉGORIE III									
	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	Modèle 7	(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)	Modèle 4	Modèle 5	Modèle 6	Modèle 7	(1, 1, 0) (0, 1, 1)	(1, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 2)	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6	Modèle 7	(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)
1%	31	19%	18	11%	29	17%	26	16%	62	37%	21	13%	20	12%	42	25%	56	34%	64	38%	73	44%	89	53%	100	60%	116	69%	121	72%
5%	45	27%	36	22%	46	28%	41	25%	82	49%	49	29%	56	34%	64	38%	73	44%	89	53%	100	60%	116	69%	121	72%				
10%	61	37%	48	29%	56	34%	55	33%	89	53%	60	36%	73	44%	89	53%	100	60%	116	69%	121	72%								
15%	72	43%	57	34%	69	41%	66	40%	101	60%	71	43%	89	53%	100	60%	116	69%	121	72%										
20%	83	50%	62	37%	80	48%	76	45%	106	64%	80	48%	95	57%	104	62%	116	69%	121	72%										
30%	100	60%	77	46%	94	56%	88	53%	119	71%	95	57%	104	62%	116	69%	121	72%												
40%	111	66%	97	58%	107	64%	99	59%	127	76%	104	62%	116	69%	121	72%														
50%	121	72%	106	63%	118	71%	113	68%	135	81%	117	70%	121	72%																
60%	131	78%	121	72%	128	77%	129	77%	141	84%	127	76%	131	78%																

En examinant la partie non saisonnière d'un modèle ARMMI, soit la partie qui explique la tendance-cycle et les valeurs extrêmes, on peut diviser les modèles en trois catégories. La catégorie I comprend les modèles 1, 2 et 3, dont la partie ordinaire renferme seulement une ou deux différences d'ordre un et un ou deux paramètres de moyenne mobile. La catégorie III comprend les modèles 5, 6 et 7, dont la partie ordinaire comporte seulement une différence d'ordre un et quelques paramètres autorégressifs. Le modèle 4 (catégorie II) constitue une catégorie en soi; sa partie non saisonnière est mixte. On peut constater que les parties saisonnières de tous les modèles sauf le modèle 7 sont identiques.

Bien que les huit critères soient analysés séparément dans cette section, plusieurs d'entre eux sont interdépendants. Par exemple, nous verrons que le nombre excessif de paramètres dans le modèle 4 cause des problèmes liés à la non-stationnarité, à la non-inversibilité, à un nombre trop grand ou trop faible de différences et à la corrélation entre les paramètres. Dans les sections 3 et 4, nous examinons les extrapolations produites à partir des sept modèles ARMMI pour des valeurs à l'intérieur des échantillons. Les modèles ont été ajustés à chaque série pour estimer les paramètres qu'on doit utiliser pour calculer les estimations relatives aux trois dernières années. C'est ainsi que les prévisions des modèles ARMMI sont évaluées par le logiciel X-11-ARMMI.

3.1 Critères concernant l'ajustement des modèles ARMMI les plus simples

La condition de stationnarité exige que toutes les racines de l'équation caractéristique d'un modèle autorégressif se trouvent à l'intérieur du cercle unité. Le tableau 1 montre que seul le modèle 4 s'avère non stationnaire, dans trois cas. Ce résultat semble attribuable au nombre excessif de paramètres dans ce modèle.

Pour qu'un modèle soit inversible, il faut que les racines de l'équation caractéristique de la moyenne mobile se trouvent à l'intérieur du cercle unité. Seul le modèle 4 produit de nombreux cas de non-inversibilité, 20% des séries, comme on peut le constater au tableau 2.

Tableau 1

Rejets dus à la non-stationnarité											
SEUIL CRITIQUE											
CATÉGORIE I			CAT. II			CATÉGORIE III					
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 2	(0, 2, 2)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0, 1, 1)	Modèle 4	(1, 1, 0)	(0, 1, 1)	Modèle 5	(2, 1, 0)	(0, 1, 1)
Modèle 1	(0, 1, 1)	(0, 1, 1)	Modèle 3	(2, 1, 2)	(0,						

Corrélation des paramètres

Une corrélation positive ou négative élevée entre les paramètres est un signe d'ambiguïté dans les valeurs estimées puisque les paramètres peuvent alors avoir différentes valeurs et absolue de quelques-uns des éléments de la matrice des corrélations des paramètres estimés est élevée, par exemple supérieure ou égale à 0,9, on peut réduire le modèle en supprimant quelques-uns des paramètres les moins significatifs.

Erreur de prévision

Il existe diverses manières de caractériser un bon ou un mauvais modèle, mais l'erreur de prévision est toujours un critère primordial. Dans cette étude, nous utilisons le pourcentage absolu moyen d'erreur pour une année de projections

$$PAME = \frac{1}{N} \sum_{t=1}^N \left| \frac{Z_{t+\ell}}{Z_t(t)} - 1 \right| \times 100\%$$

où t est 12 ou 4 et $Z_t(t)$ est la projection pour la période d'avance t .

3. EVALUATION DES MODELES ARMMI

Les huit critères ont été répartis en deux groupes. Le premier groupe a trait au bon ajustement des modèles les plus simples, le deuxième à la qualité des prévisions. Cette distinction entre l'ajustement d'un modèle et le calcul de prévisions est importante; un modèle bien ajusté et des prévisions de bonne qualité ne sont pas équivalents.

Ces critères ont été utilisés pour évaluer et classer sept des modèles ARMMI les plus souvent appliqués, à savoir:

1. (0, 1, 1) (0, 1, 1)^s
2. (0, 1, 2) (0, 1, 1)^s
3. (0, 2, 2) (0, 1, 1)^s
4. (2, 1, 2) (0, 1, 1)^s
5. (1, 1, 0) (0, 1, 1)^s
6. (2, 1, 0) (0, 1, 1)^s
7. (2, 1, 0) (0, 1, 2)^s

où s est 12 si une série chronologique est mensuelle et 4 si elle est trimestrielle.

Ces modèles ont été ajustés à un échantillon de 167 séries chronologiques saisonnières mensuelles choisies au hasard dans onze domaines de l'économie canadienne: le système des comptes nationaux; le marché du travail; les prix; les industries manufacturières; les combustibles, l'énergie et l'exploitation minière; la construction; l'alimentation et l'agriculture; le commerce intérieur; le commerce extérieur; les transports et les finances. Une quarantaine de séries chronologiques trimestrielles relatives aux comptes nationaux et aux finances ont également été choisies pour cette expérience.

Ces séries sont, pour la plupart, de nature multiplicative, selon l'essai fondé sur le modèle de Bell Canada (Higginson, 1976), c'est-à-dire qu'on doit multiplier les différentes composantes (tendance-cycle, éléments saisonniers et variations irrégulières) pour produire les valeurs observées. Par conséquent, l'amplification de la composante saisonnière est souvent élevée, plus le niveau de la tendance est haut. Une transformation logarithmique a été appliquée aux séries multiplicatives avant l'ajustement des trois premiers et des trois derniers modèles. Le quatrième modèle a été ajusté à des séries non transformées dans tous les cas.

et $\phi(B)$ tend vers

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_{p-1} B^{p-1})(1 - B).$$

Par conséquent, il peut falloir inclure un opérateur de différence dans ce système, et le modèle AR (p) devient un modèle $ARI(p - 1, 1)$. En outre, quand λ_k tend vers 1, le système peut également devenir non stationnaire.

Nombre excessif de différences

Prenons le modèle général ARMMI (p, d, q) (P, D, Q),

$$\phi(B)\Phi(B)(1 - B)^d(1 - B^p)^pZ_t = \theta(B)\Theta(B)a_t.$$

Si des racines v_i de l'équation caractéristique $\theta(B) = 0$ tendent vers 1, c'est-à-dire si une ou plusieurs valeurs de $(1 - v_i B)$ tendent vers $(1 - B)$, on peut éliminer $(1 - B)$ des deux membres de l'équation.

Test du caractère aléatoire des a_t

L'existence d'une corrélation entre les résidus n'est pas souhaitable parce que nous voulons une estimation non biaisée des paramètres d'un processus.

Nous utilisons la variable

$$\tilde{Q} = n(n + 2) \sum_{k=1}^m (n - k)^{-1} \tilde{q}_k^2$$

qui a été proposée par Prothero et Wallis (1976) et Ljung et Box (1978) et qui constitue une forme modifiée du test khi-carré de Box et Pierce.

Dans cette formule, n est la taille de l'échantillon, $k = 1, 2, \dots, m$, sont les divers retards et les \tilde{q}_k sont les coefficients d'autocorrélation. La variable \tilde{Q} est utilisée pour vérifier si les résidus sont aléatoires.

Paramètres non significatifs

En général, quand on augmente le nombre de paramètres dans un modèle, la moyenne de la somme des carrés, σ_a^2 diminue. Mais seuls les paramètres dont la valeur est élevée, ou ceux dont la valeur est significativement différente de 0, peuvent provoquer une baisse notable de σ_a^2 . Pour savoir si la valeur d'un paramètre est significative, on peut recourir à un test F (Pandit et Wu 1983):

$$F = \frac{A_1 - A_0}{s} \div \frac{A_0}{N - r} \sim F(s, N - r)$$

où r est le nombre de paramètres dans le modèle et s est le nombre de paramètres limités à zéro. N est le nombre d'observations, A_0 est la valeur la moins élevée de la somme des carrés du modèle limite et A_1 est la valeur la plus élevée de la somme des carrés du modèle limite.

Toutefois, dans la présente étude, nous employons deux constantes, 0.05 et 0.10, comme seuils critiques pour déterminer si un paramètre est significatif.

2. LES CRITÈRES

Dans cette section, nous décrivons brièvement les huit critères d'évaluation des modèles.

Stabilité

Un processus Z_t est soit stationnaire, soit non stationnaire. Le degré de stabilité indique à quel point la "mémoire" du système retient les perturbations passées, a_{t-j} , $j = 1, 2, \dots$, et à quel rythme l'effet d'une perturbation sur le système est dissipé. Pour un processus

$$Z_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$$

$$= \psi(B)a_t,$$

où $a_t \sim NID(0, \sigma_a^2)$, le filtre est considéré comme stable si la série $\{\psi_j\}$ est convergente. Un modèle ARMMI (p, d, q),

$$\phi(B)(1 - B)^d Z_t = \theta(B)a_t,$$

est stable si toutes les racines λ_j de l'équation caractéristique

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p = (1 - \lambda_1 B)(1 - \lambda_2 B) \dots (1 - \lambda_p B) = 0$$

se trouvent strictement à l'intérieur du cercle unité, c'est-à-dire si $|\lambda_j| < 1$.

Inversibilité

Le processus Z_t peut être exprimé sous la forme suivante:

$$Z_t = a_t + \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \dots$$

Le système est inversible si la série $\{\pi_j\}$ est convergente. Ce critère est considéré comme primordial parce que, si le système n'est pas inversible, la fonction génératrice $\pi(B)$ des π_j s croît sans limite. Dans ce cas, l'état présent du système dépendrait d'événements situés dans le passé infini et le processus n'aurait aucun sens concret.

Un modèle général ARMMI (p, d, q), est inversible si les racines v_j de l'équation caractéristique

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q = (1 - v_1 B)(1 - v_2 B) \dots (1 - v_q B) = 0$$

se trouvent strictement à l'intérieur du cercle unité, c'est-à-dire si $|v_j| < 1$.

Nombre insuffisant de différences

Dans le modèle AR(p), quand une ou plusieurs racines caractéristiques λ_j , supposons qu'il s'agit de λ_k tendent vers 1, on peut écrire

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

$$= (1 - \lambda_1 B) \dots (1 - \lambda_{k-1} B)(1 - \lambda_k B) \dots (1 - \lambda_p B)$$

$$= (1 - \lambda_1 B) \dots (1 - \lambda_{k-1} B)(1 - \lambda_{k+1} B) \dots (1 - \lambda_p B)(1 - \lambda_k B),$$

Évaluation de modèles ARMMI appliqués à des séries chronologiques¹

KIM CHIU, JOHN HIGGINSON et GUY HUOT²

RÉSUMÉ

Cette étude porte sur l'évaluation des prévisions calculées à partir de quelques-uns des modèles ARMMI (modèles autorégressifs à moyenne mobile intégrée) les plus utilisés. Ces modèles ont été ajustés à un échantillon de deux cents séries chronologiques saisonnières relatives à onze domaines de l'économie canadienne. Les modèles ont été évalués en fonction de huit critères: l'erreur moyenne de prévision pour les trois dernières années, la variabilité du khi-caré pour le test du caractère aléatoire des résidus, la détection de paramètres non significatifs, d'un trop grand nombre de différences, d'un nombre insuffisant de différences, d'une corrélation entre les paramètres, la stationnarité et l'inversibilité. Les modèles sont comparés à l'aide de classements globaux et conditionnels le tout étant illustré à l'aide de graphiques.

MOTS CLÉS: X-11-ARMMI; classement global; classement conditionnel; critères

1. INTRODUCTION

Notre environnement socio-économique est instable et incertain; l'inflation, les récessions et la dégradation du milieu par la pollution sont quelques-uns des facteurs qui contribuent à accroître l'instabilité. Pour essayer de résoudre ce problème, nous utilisons une méthode de prévision qui permet d'évaluer les effets des changements fréquents qui surviennent. Les modèles ARMMI (Box et Jenkins 1970) sont assez souples pour l'analyse de ces changements dans les séries chronologiques.

Cet article a pour objet d'examiner huit critères applicables à la méthode de Box-Jenkins pour évaluer la qualité de l'ajustement et des prévisions des modèles les plus souvent utilisés pour l'analyse des séries chronologiques économiques au Canada. Déterminer les modèles qui produisent les meilleurs résultats est une question importante pour des programmes, comme le logiciel X-11-ARMMI (Dagum 1980), qui ajustent automatiquement un petit groupe fixe de modèles (trois modèles dans le cas du X-11-ARMMI) aux séries chronologiques.

La section 2 présente huit critères: l'erreur moyenne de prévision pour les trois dernières années; la variabilité du khi-caré pour le test du caractère aléatoire des résidus; la détection de paramètres non significatifs, d'un trop grand nombre de différences, d'un nombre insuffisant de différences et d'une corrélation entre les paramètres; la stationnarité et l'inversibilité. La section 3 fournit une description de ces critères et un résumé des résultats. La section 4 présente des classements globaux et conditionnels des modèles, et dans la section 5, on compare les valeurs extrapolées à l'intérieur et à l'extérieur des échantillons pour les trois dernières années.

¹ Exposé présenté (1) à la conférence sur les prévisions commerciales et économiques dans le cadre du symposium canadien sur la recherche opérationnelle, Ottawa, mai 1984 et (2) à la section sur la statistique commerciale et économique dans le cadre des réunions de l'American Statistical Association, Philadelphie, août 1984.

² K. Chiu, J. Higginson et G. Huot, Division de la recherche et de l'analyse des chroniques, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6.

- DAHMSSTRÖM, P., et HAGNELL, M. (1975). Multivariate stratification of primary sampling units in multi-stage sampling with an application to SCB's general purpose sample. Document de recherche, Université de Lund.
- DREW, J.D., BÉLANGER, Y., FOY, P. (1985). Multivariate clustering algorithm for stratification and its application to the Canadian Labour Force Survey. Document technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada (en voie de rédaction).
- FELLEG, I.P., GRAY, G.B., et PLATEK, R. (1967). The new design of the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 62, 421-453.
- FRIEDMAN, H.P., et RUBIN, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.
- GRAY, G.B. (1971). Joint probability of selection of units in systematic samples. *Proceedings of American Statistical Association*, 271-276.
- HARTLEY, H.O., et RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annales de la statistique mathématique*, 33, 350-374.
- HIDIROGLOU, M.A., et GRAY, G.B. (1980). Construction of joint probability of selection for systematic PPS sampling. *Journal of the Royal Statistical Society*, C29, 107-112.
- HORVITZ, D.G., et THOMSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- LEMAITRE, G. (1983). Some results from Time and Cost Study. Document technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- PLATEK, R., et SINGH, M.P. (1976). *Méthodologie de l'enquête sur la population active du Canada*. N° 71-526 au catalogue, Statistique Canada.
- RAO, J.N.K., HARTLEY, H.O., et COCHRAN, W.G. (1962). A simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, B24, 482-491.
- SINGH, M.P., et DREW, J.D. (1981). Research plans for the redesign of the Canadian Labour Force Survey. *Proceedings of the Section of Survey Research Methods, American Statistical Association Meetings*.
- SINGH, M.P., DREW, J.D., et CHOUDHRY, G.H. (1984). Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981. *Techniques d'enquête*, 10, 139-154.

i) Algorithme de transfert de Friedman-Rubin (1967)

Cet algorithme non hiérarchique, qui a été choisi pour la stratification de l'échantillon de l'EPA (Drew et coll. 1985), répartit les unités en ensembles aléatoires et recherche un optimum local en déplaçant une unité à la fois d'un ensemble à un autre si ce transfert réduit la valeur de *M*. Avant chaque déplacement, l'algorithme vérifie également si les contraintes relatives à la taille des ensembles seraient respectées. Une approximation de l'optimum global est atteinte à partir de plusieurs points d'origine choisis au hasard. L'algorithme de Friedman-Rubin présente un inconvénient dans ce cas. Etant donné que les contraintes relatives à la taille des ensembles sont strictes parce que les tâches des interviewers doivent être à peu près égales, le déplacement des unités entre les différents ensembles est limité.

ii) Algorithme d'échange de Dahmström-Hagnell (1975)

Cet algorithme est semblable à l'algorithme de Friedman-Rubin, sauf qu'il repose sur l'échange d'une unité contre une autre entre deux ensembles et non sur le transfert d'une seule unité à la fois. Cet algorithme est donc mieux adapté aux cas où les contraintes relatives à la taille des ensembles sont strictes.

iii) Algorithmes mixtes

Définissons un cycle d'un algorithme mixte comme l'application de l'algorithme d'échange suivie de l'application de l'algorithme de transfert. Nous avons examiné des algorithmes mixtes à un et à deux cycles.

L'algorithme mixte à deux cycles fonctionne mieux que les autres. Il requiert le nombre le plus faible de points d'origine aléatoires et le moins de frais de calcul pour atteindre le même degré d'optimalité que les autres algorithmes. Les résultats de l'essai des algorithmes mixtes à un et à deux cycles pour 21 échantillons sont résumés dans le tableau suivant.

		Un cycle		Deux cycles	
		Nombre d'origines aléatoires		Nombre d'origines aléatoires	
		1	2	4	2
M*		336.18	329.19	325.65	325.51
Ecart-type		15.84	15.45	15.67	15.69
Frais de calcul (\$)		5.94	11.24	21.67	15.12
* Moyenne pour 21 échantillons.					

BIBLIOGRAPHIE

CONNOR, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

La variance de cet estimateur est

$$V(\hat{Y}^{HT}) = \sum_{i=1}^N \sum_{j<i}^N (\pi_i \pi_j - \pi_{ij})^2 \left(\frac{\pi_i}{Y_i} - \frac{\pi_j}{Y_j} \right)^2,$$

où π_{ij} est la probabilité que les $i^{\text{ème}}$ et $j^{\text{ème}}$ unités soient comprises dans S . Hartley et Rao (1962) ont proposé une formule asymptotique pour les π_{ij} 's.

Il existe également une formule exacte mise au point par Connor (1966), mais elle est assez complexe. Récemment, Hidiroglou et Gray (1980) ont élaboré un algorithme informatique à l'aide d'une forme modifiée de la formule de Connor qui a été établie par Gray (1971); cette formule modifiée a été utilisée dans notre étude et comparée avec l'approximation de Hartley-Rao. On a constaté que les approximations de Hartley-Rao sont très proches des valeurs exactes pour $N \geq 16$. Nous avons décidé d'utiliser l'algorithme de Hidiroglou-Gray pour $N < 16$ et l'approximation de Hartley-Rao pour $N \geq 16$ puisque le nombre de calculs dans l'algorithme croît exponentiellement à mesure que N grandit.

ANNEXE B

Simulation du coût de P_2 et P_1

Pour estimer r , le ratio entre les paiements et les frais pour les déplacements des interviewers entre leur domicile et le secteur où ils travaillent, entre les UPE et entre les unités secondaires dans les plans de sondage P_2 et P_1 pour les régions NAR, on a procédé à une étude de Monte Carlo. Les bases de sondage de P_1 et P_2 ont été simulées jusqu'au niveau des unités secondaires à l'aide des données du recensement pour chacune des 11 RE définies pour cette étude. Cinquante échantillons ont été tirés pour chaque plan de sondage et les unités secondaires choisies pour chaque échantillon ont été regroupées en tâches d'interviewers d'une manière géographiquement optimale. Si $\bar{M}_{(1)}$ et $\bar{M}_{(2)}$ sont les mesures moyennes de la dispersion géographique à l'intérieur des tâches définies pour les plans P_1 et P_2 , la valeur estimée de r est

$$\bar{M}_{(2)}/\bar{M}_{(1)}.$$

La variable M pour un échantillon donné a été définie de la manière suivante. Supposons que k interviewers sont affectés à une RE et que $G_i = \{U_{ij}; j = 1, 2, \dots, n_i\}$ le cent- $i^{\text{ème}}$ interviewer et comprend n_i unités secondaires d'échantillonnage. Soit (x_{ij}, y_{ij}) le cent- $i^{\text{ème}}$ de la population de U_{ij} défini en coordonnées euclidiennes. La valeur de M pour cette RE est

$$M = \sum_{k=1}^K M_k = \sum_{i=1}^f \sum_{n_i} \{(x_{ij} - x_j)^2 + (y_{ij} - y_j)^2\}^{1/2},$$

où (x_j, y_j) est le centre de G_i , c'est-à-dire $x_i = 1/n_i \sum_{j=1}^{n_i} x_{ij}$; $y_i = 1/n_i \sum_{j=1}^{n_i} y_{ij}$.

La définition des tâches optimales pour les interviewers, c'est-à-dire la minimisation de la variable M , se réduit à un problème de classification. Les algorithmes de classification suivants ont été examinés:

Tableau 7
Efficacité relative de l'échantillon remanié par rapport à l'ancien échantillon pour l'estimation du nombre de chômeurs

Province	Ratio des coûts* ($= \frac{C(A)}{C(M)}$)	Ratio des variances ($= \frac{V(A)}{V(M)}$)	Efficacité relative ($= \frac{C(M)V(M)}{C(A)V(A)}$)
----------	---	--	--

Terre-Neuve	1.19	1.00	1.19
Ile-du-Prince-Edouard	1.10	1.13	1.24
Nouvelle-Ecosse	1.22	1.04	1.27
Nouveau-Brunswick	1.17	0.99	1.16
Québec	1.15	0.95	1.09
Ontario	1.13	1.03	1.16
Manitoba	1.17	0.96	1.12
Saskatchewan	1.23	1.02	1.25
Alberta**	1.15	1.00	1.15
Colombie-Britannique	1.15	1.01	1.16
Canada	1.17	0.99	1.16

* Pour le nouvel échantillon, les interviews sont menées par téléphone dans les régions NAR à partir du deuxième mois qu'un ménage fait partie de l'échantillon, alors qu'elles étaient menées sur place dans le cas de l'ancien échantillon.

** Echantillon supplémentaire non inclus.

cien échantillon et celui du nouvel échantillon est égal à 1.16 (tableau 7). L'échantillon de l'Alberta a par la suite été élargi de 1,300 ménages pour produire des données sur un nombre accru de régions intraprovinciales et sa répartition a été modifiée. En tout, l'échantillon actuel de l'EPA compte 52,900 ménages par mois.

ANNEXE A

Formule et méthode de calcul de la variance pour l'ESCAPPT

Supposons qu'un échantillon de taille n est choisi par la méthode d'échantillonnage systématique avec classement aléatoire et probabilité proportionnelle à la taille (ESCAPPT) dans une population de N unités. Soit p_i la mesure normalisée de la taille de la $i^{\text{ème}}$ unité, de sorte que $\sum_{i=1}^N p_i = 1$. L'estimateur de Horvitz-Thomson pour le total Y d'une caractéristique y est (Horvitz et Thomson 1952):

$$Y_{HT} = \sum_{i \in S} \frac{y_i}{\pi_i},$$

où S = l'échantillon de taille n

y_i = la valeur de y pour la $i^{\text{ème}}$ unité,

$\pi_i = np_i$, la probabilité que la $i^{\text{ème}}$ unité fasse partie de S .

Tableau 6
Efficacité relative de l'échantillon remanié par rapport à l'ancien échantillon, selon la province (nombre de chômeurs)

Province	Ratio des coûts ($= \frac{C(A)}{C(M)}$)	Ratio des variances ($= \frac{V(A)}{V(M)}$)	Efficacité relative ($= \frac{C(M)V(A)}{C(A)V(M)}$)
Terre-Neuve	1.00	1.00	1.00
Ile-du-Prince-Edouard	1.01	1.02	1.03
Nouvelle-Ecosse	1.04	1.14	1.18
Nouveau- Brunswick	1.01	0.98	0.99
Québec	1.03	1.06	1.09
Ontario	1.04	1.08	1.12
Manitoba	1.01	1.03	1.04
Saskatchewan	1.05	1.06	1.12
Alberta	1.01	1.01	1.02
Colombie-Britannique	1.02	1.09	1.11
Canada	1.03	1.07	1.10

5. CONCLUSIONS

Les modifications suivantes ont été apportées au plan de sondage de l'EPA à cause des résultats des études sur les moyens d'optimiser le coût et la variance: élimination d'une étape de l'échantillonnage dans les parties rurales des régions NAR, choix d'un plan de sondage comportant une stratification des régions rurales et urbaines, choix d'un plan de sondage à deux degrés pour les régions NAR de l'Ile-du-Prince-Edouard, augmentation du nombre de strates dans les régions NAR à cause de la sélection de seulement 2 ou 3 UPE par strate et application d'une nouvelle répartition optimale de l'échantillon entre les régions NAR et AR. Les autres paramètres du plan de sondage que Fellegi, Gray et Platek (1967) ont trouvés presque optimaux le demeurent, par exemple le nombre de logements à choisir par UPE dans les régions AR.

Les améliorations sur le plan de l'efficacité ont permis de réduire de 7% la taille globale de l'échantillon de l'EPA et d'assurer le degré nécessaire de fiabilité des données intraprovinciales (Singh et coll., 1984) sans nuire à la fiabilité des estimations provinciales et nationales. Les seules exceptions sont les provinces de Québec et de l'Alberta, où les critères de fiabilité des données intraprovinciales sont plus stricts qu'ailleurs et entraînent une légère baisse dans la fiabilité des données provinciales. Le tableau 7 présente les ratios entre le coût, la variance et le produit du coût et de la variance de l'ancien échantillon (fondé sur l'ancien plan de sondage avec 55,000 ménages par mois et aucune interview téléphonique dans les régions NAR) et de l'échantillon remanié (nouveau plan de sondage avec 51,600 ménages par mois et des interviews téléphoniques dans les régions NAR). Les fortes baisses de coût sont dues à l'exécution des interviews par téléphone du deuxième au sixième mois qu'un ménage fait partie de l'échantillon dans les régions NAR et à la réduction de la taille globale de l'échantillon. Pour l'ensemble du Canada, le ratio entre le produit du coût et de la variance de l'an-

Tableau 5
Pourcentage de l'échantillon dans les régions AR selon la province pour
1) l'ancien échantillon, 2) une répartition proportionnelle,
3) la répartition optimale et 4) l'échantillon remanié

Province	Ancien échantillon	Répartition proportionnelle	Répartition optimale	Echantillon remanié
Terre-Neuve	41.8	51.3	42.6	44.6
Ile-du-Prince-Edouard	26.6	32.8	32.8	28.9
Nouvelle-Ecosse	37.3	57.4	58.8	51.9
Nouveau-Brunswick	49.5	52.5	47.4	53.6
Québec	56.8	74.8	71.6	68.9
Ontario	62.5	79.1	78.8	75.0
Manitoba	54.1	71.0	76.4	56.4
Saskatchewan	44.7	51.8	62.1	56.8
Alberta	60.0	68.6	72.6	62.3
Colombie-Britannique	58.0	78.0	74.6	69.7
Canada	53.2	67.1	67.4	62.3

Le tableau 5 indique le pourcentage de l'échantillon dans les régions AR selon les réparti-
tions suivantes: i) l'ancien plan de sondage, ii) une répartition proportionnelle, iii) la réparti-
tion optimale découlant du modèle du coût et de variance proposé et iv) la répartition choisie
pour le nouvel échantillon de l'EPA. Il n'a pas été possible d'opter pour la répartition op-
timale à cause de contraintes relatives à la fiabilité des données infraprovinciales. Dans la
plupart des cas, les différences entre la répartition optimale et la répartition choisie sont petites.
On a constaté que la répartition optimale ressemble de près à une répartition proportion-
nelle et diffère beaucoup de la répartition imposée par l'ancien plan de sondage.

Le tableau 6 présente des chiffres sur les améliorations découlant uniquement d'une nouvelle
répartition sans modification des tailles (celles de l'ancien plan de sondage) des échantillons
provinciaux et avec des fractions de sondage uniformes à l'intérieur des deux types de régions.
Pour ce tableau, on a utilisé les coûts et les variances unitaires décrits plus haut pour calculer
le coût total et la variance totale, $C^{(A)}$, $C^{(M)}$, $V^{(A)}$, $V^{(M)}$, pour l'ancien et le nouveau plan de
sondage respectivement. La nouvelle répartition entraînerait une baisse de 3% dans le coût
total et de 7% dans la variance totale du nombre de chômeurs, et l'efficacité relative du point
de vue du produit du coût et de la variance (définie au tableau 6) serait égale à 1.10. N'aurait
été des normes relatives aux données infraprovinciales, la répartition optimale aurait permis
d'atteindre une efficacité relative de 1.12.

L'efficacité relative de l'échantillon remanié par rapport à l'ancien échantillon est examinée
à la section suivante.

4. OPTIMISATION DU COÛT ET DE LA VARIANCE EN FONCTION DE LA RÉPARTITION DE L'ÉCHANTILLON ENTRE LES RÉGIONS AR ET NAR

L'étape suivante de l'optimisation du coût et de la variance dans le plan de sondage de l'EPA était l'optimisation de la répartition de l'échantillon entre les régions AR et NAR. Nous avons utilisé les modèles simples de coût et de variance qui ont été présentés par Fellegi, Gray, et Platek, (1967):

$$(4.1) \quad \text{coût:} \quad C = \sum_{j=1}^2 C_j \frac{W_j}{P_j},$$

$$(4.2) \quad \text{variance:} \quad V = \sum_{j=1}^2 W_j P_j \sigma_j^2,$$

où j = type de région (= 1 région AR; = 2 région NAR),
 C_j = coût unitaire (par personne),
 P_j = population,
 $1/W_j$ = fraction de sondage,
 σ_j^2 = variance unitaire.

Fellegi et coll. ont montré que si l'on minimise C pour une valeur fixe de V , le ratio des fractions de sondage est

$$(4.3) \quad \frac{W_1}{W_2} = \frac{\sigma_1}{\sigma_2} \left(\frac{C_1}{C_2} \right)^{1/2}$$

Avec les autres critères d'optimisation décrits à la section 1, ce ratio demeure le même. Les paramètres ont été estimés de la manière suivante:

- (i) **Coût unitaire:** On a utilisé des données recueillies dans le passé sur le coût par logement dans les deux types de régions. Ces chiffres ont été réduits de 10% dans le cas des régions NAR pour tenir compte de l'effet estimé de l'utilisation d'interviews téléphoniques pour tous les groupes de renouvellement sauf celui des ménages qui sont à leur premier mois d'inclusion dans l'échantillon remanié.
- (ii) **Variance unitaire:** La répartition de l'échantillon a été optimisée par rapport au nombre de chômeurs, dont la variance est:

$$(4.4) \quad \sigma_j^2 = \beta_j \frac{P_j}{n_j} \left(1 - \frac{P_j}{n_j} \right); j = 1, 2$$

où β_j est l'effet du plan de sondage sur l'estimation du nombre de chômeurs et n_j , le nombre de chômeurs.

On a utilisé des données passées sur l'effet du plan dans les deux types de régions et on les a réduites pour tenir compte des améliorations décrites aux sections 2 et 3 concernant la structure du plan de sondage des régions NAR et AR. Les valeurs du nombre de chômeurs reposent sur les moyennes des données de l'EPA pour la période 1980-1982, ce qui semble juste étant donné que les prévisions à moyen terme n'annoncent pas un retour aux nombres de chômeurs observés avant la récession de 1982, et les chiffres de population proviennent du recensement de 1981.

Tableau 4
Efficacité relative de P_1 par rapport à P_2 , du point de vue du coût et de la variance

ER	Personnes occupées		Personnes occupées	
	Chômeurs		Chômeurs	
Efficacité statistique	V_{P_1}/V_{P_2}		Efficacité	
	sur le plan		du coût	
Efficacité relative du point de vue du produit	C_P^T/C_P^T		$V_{P_1}C_{P_1}^T/V_{P_2}C_{P_2}^T$	
	du coût et de la variance		du coût et de la variance	

22	1.09	0.93	1.02	1.11	0.95
32	0.91	0.72	1.03	0.94	0.74
41	1.14	0.86	1.23	1.40	1.06
44	1.39	1.14	1.19	1.65	1.37
51	0.96	1.01	1.03	0.99	1.04
56	1.12	1.51	1.10	1.23	1.66
63	1.35	1.06	1.05	1.41	1.11
72	1.00	0.91	1.06	1.06	0.96
82	1.09	1.01	1.18	1.27	1.19
86	1.20	1.05	1.04	1.25	1.09
96	1.38	1.05	1.07	1.48	1.12
Total*	1.16	0.97	1.08	1.25	1.05

* Moyenne pondérée en fonction de la taille de la population.

Les résultats d'une analyse du coût et de la variance ont montré que l'efficacité statistique relative de P_3 par rapport à P_1 était égale à 2.39 pour le nombre de personnes occupées et à 1.20 pour le nombre de chômeurs, alors que le coût du plan P_3 était supérieur de seulement 8% au coût de P_2 . Ainsi, étant donné que, du point de vue du produit du coût et de la variance, l'efficacité relative globale de P_3 par rapport à P_2 était de 2.21 pour l'estimation du nombre de personnes occupées et de 1.11 dans le cas du nombre de chômeurs, on a choisi le plan P_3 .

3.6 Nombre d'UPÉ choisies par strate

Dans les plans P_1 et P_2 , un rendement de 55 à 60 logements par UPÉ a été fixé pour l'échantillon, ce qui correspond à la tâche d'un interviewer. Dans près de la moitié des RE, la taille de l'échantillon permettait de choisir seulement 2 ou 3 UPÉ. On a décidé qu'il serait injustifié de stratifier davantage ces RE parce qu'il devrait y avoir au moins 2 UPÉ par strate pour qu'on puisse estimer la variance sans biais.

Pour les autres RE, on a songé à des strates de 4 ou 5 UPÉ, ce qui offrirait plus de souplesse pour réduire la taille de l'échantillon dans une région si, par exemple, une partie des répondants de cet échantillon étaient ultérieurement interviewés par téléphone à l'intérieur d'une base de sondage double. On a cependant décidé que chaque strate comprendrait 2 ou 3 UPÉ parce qu'on peut ainsi réduire de 14.8% la variance de l'estimation du nombre de personnes occupées et de 5.4% la variance de l'estimation du nombre de chômeurs dans ces RE. Une description détaillée de la méthode de stratification adoptée a été présentée par Drew, Bélanger, et Foy (1985).

Tableau 3

Valeurs des paramètres du modèle de coût pour les régions NAR et coût relatif de P_1 par rapport à P_2 dans une enquête où les interviews sont menées par téléphone

RE	F_0	F_1	F_2	E_1	E_2	α	r	$C_T^{P_1}$	$C_T^{P_2}$	$C_T^{P_1}/C_T^{P_2}$
96	2.03	0.81	1.22	0.75	0.85	0.84	0.75	5.07	4.74	1.07
86	1.88	1.12	1.01	1.20	0.94	0.86	0.90	5.55	5.35	1.04
82	1.88	1.12	1.01	1.20	0.94	0.86	0.57	5.55	4.69	1.18
72	1.92	0.96	1.13	1.05	1.09	0.85	0.82	5.52	5.21	1.06
63	2.07	1.03	1.03	1.19	0.97	0.75	0.87	5.66	5.41	1.05
56	1.94	0.80	1.07	0.81	0.75	0.84	0.68	4.82	4.39	1.10
51	1.94	0.80	1.07	0.81	0.75	0.84	0.89	4.82	4.67	1.03
44	2.04	0.94	0.94	0.96	0.69	0.84	0.50	5.01	4.21	1.19
41	2.04	0.94	0.94	0.96	0.69	0.84	0.42	5.01	4.08	1.23
32	2.13	0.86	1.11	0.90	0.97	0.84	0.88	5.35	5.17	1.03
22	2.05	0.74	1.31	0.95	0.92	0.85	0.93	5.38	5.28	1.02

Chose peu surprenante, les coûts des déplacements des interviewers entre les UPB et entre les unités secondaires sont plus élevés dans le plan P_1 à cause du manque fréquent de con-
tiguïté entre les parties rurales et urbaines des UPB. Le coefficient moyen de réduction dans
ces coûts dans le plan P_2 a été estimé à partir des données utilisées pour le tableau 3 et il
s'est avéré que le coût relatif de P_1 par rapport à P_2 pour l'ensemble des RE est de 1.08
(tableau 4).

Analyse du coût et de la variance: Comparaison de P_2 à P_1

Le tableau 4 présente les valeurs relatives du produit du coût et de la variance du plan
 P_2 par rapport à P_1 pour une enquête dans laquelle les interviews sont menées par téléphone.
Globalement, du point de vue du produit du coût et de la variance, le plan P_2 est plus ef-
ficace que P_1 de 25% et de 5% pour les estimations du nombre de personnes occupées et
du nombre de chômeurs respectivement.

À la lumière de ces résultats, il a été décidé que le plan P_2 serait utilisé pour les 2/3 des
RE dans lesquelles il est possible de former au moins une strate urbaine et une strate rurale
et que le plan P_1 serait utilisé dans les autres cas.

3.5 Plan de sondage spécial à deux degrés pour l'Île-du-Prince-Édouard

Dans la plus petite province du Canada, l'Île-du-Prince-Édouard, où des fractions de son-
dage de 4% sont nécessaires pour recueillir des données fiables, le plan de sondage P_3 , fondé
sur un échantillon stratifié de SD et de logements, a été proposé pour remplacer le plan P_2 .
Dans le plan P_3 , l'échantillon n'est pas divisé en UPB géographiquement contiguës qui
correspondent aux tâches des interviewers parce qu'on a supposé que, puisque les fractions
de sondage sont élevées, l'augmentation des coûts de la collecte des données serait plus que
compensée par des baisses de la variance dues à l'élimination d'une étape de l'échantillon-
nage et aux améliorations découlant d'un accroissement du nombre de strates (P_3 comprend
jusqu'à 4 fois le nombre de strates compris dans l'échantillon fondé sur P_2).

Tableau 2

Proportion de la variance totale due à chaque étape de l'échantillonnage dans le plan de sondage actuel et pourcentage de réduction dans la variance totale attribuable à l'élimination des grappes dans les régions rurales; $100 (1 - \frac{V_{P_1}}{V_{P_0}})$

Caractéristique	Pourcentage de la variance totale dû à chaque étape				Pourcentage de réduction de la variance;			
	Régions urbaines		Régions rurales		$100 (1 - \frac{V_{P_1}}{V_{P_0}})$			
	1 ^{ère} étape	2 ^e étape	3 ^e étape	4 ^e étape				
	étape	étape	étape	étape				
Population active	14.5	12.9	10.8	5.8	40.5	15.5		30.5
Nombre de personnes occupées	21.2	11.2	10.4	6.3	35.0	15.8		27.1
Nombre de chômeurs	12.6	15.8	16.6	4.8	33.0	17.2		24.8
Nombre d'inactifs	24.7	11.9	10.7	4.8	32.9	15.1		22.9
Nombre de personnes occupées dans le secteur agricole	42.4	1.0	0.8	12.3	30.8	12.6		20.4
Nombre de personnes occupées en dehors du secteur agricole	23.3	12.7	11.9	5.6	31.7	14.8		21.8

En réalité, les gains d'efficacité ne seraient peut-être pas si élevés puisque les variables estimées et les mesures de la taille de la population se rapportent à la même période, ce qui n'arrive pas en pratique. On n'a toutefois pas essayé de dégonfler ces chiffres, étant donné que le choix entre P_1 et P_0 est clair, du point de vue tant de la variance que du déroulement des opérations de collecte (voir section 3.1). On s'est donc penché sur la question du choix entre P_1 et P_2 .

Analyse de la variance: Comparaison de P_2 à P_1

Pour cette analyse, le nombre de RE a été porté à 11 et les estimations des variables visées (nombre de personnes occupées et nombre de chômeurs) reposent sur les données du recensement de 1976 tandis que les mesures de la taille de la population reposent sur les données du recensement de 1971. En outre, les variances ont été calculées pour des estimations par le quotient utilisant l'ensemble de la population.

Pour l'ensemble des RE, l'efficacité statistique relative du plan P_2 par rapport au plan P_1 était de 1.16 pour l'estimation du nombre de personnes occupées et de 0.97 pour l'estimation du nombre de chômeurs (tableau 4).

Analyse du coût: Comparaison de P_2 à P_1

Les valeurs de tous les paramètres du modèle de coût sont présentées au tableau 3, qui indique également les résultats obtenus pour $C_{P_1}^T$, $C_{P_2}^T$ et le ratio de ces deux coûts.

Le modèle de coût élaboré pour le plan de sondage P_1 avec toutes les interviews ayant lieu sur place est le suivant:

$$C_{P_1} = F_0 + F_1 + F_2 + E_1 + E_2$$

où F_0 = paiement fixe pour les interviews,
 F_1 = paiement pour les déplacements des interviewers entre leur domicile et le secteur où ils travaillent, entre les UPB et entre les unités secondaires,
 F_2 = paiement pour les déplacements à l'intérieur des unités secondaires (d'un logement à un autre),
 E_1 = frais remboursés pour les déplacements des interviewers entre leur domicile et le secteur où ils travaillent, entre les unités secondaires,
 E_2 = frais remboursés pour les déplacements d'un logement à un autre.

Les paiements représentent la rémunération versée aux interviewers pour le temps qu'ils consacrent à leur travail et les frais se rapportent à la distance parcourue. Tous les paramètres sont exprimés sous forme de coût par logement.

Si les interviews sont menées par téléphone, le modèle est le suivant

$$C_{P_1}^T = F_0 + \alpha(F_1 + F_2 + E_1 + E_2),$$

où α est le facteur par lequel les temps et les distances de déplacement diminueraient à cause de l'emploi du téléphone.

Or si l'on suppose que le plan P_2 modifierait les valeurs de F_1 et E_1 , disons par un facteur r , mais non les autres composantes, le modèle de coût est le suivant:

$$C_{P_2}^T = F_0 + \alpha r(F_1 + E_1) + \alpha(F_2 + E_2).$$

Les paramètres de $C_{P_1}^T$ et $C_{P_2}^T$ ont été estimés de la manière suivante:

F_0, F_1, F_2, E_1, E_2 : Pour P_0 , ces paramètres ont été estimés à partir des données recueillies au cours d'une étude des temps et des coûts (Lemaître 1983) dans le cadre du programme de recherche en vue du remaniement du plan de sondage de l'EPA. Comme la mise à l'essai de P_1 n'a révélé aucune différence notable entre les coûts de la collecte des données pour P_0 et P_1 , on a supposé que ces paramètres ont les mêmes valeurs pour P_0 et P_1 .

α : Les essais d'interviews téléphoniques menés dans le cadre du programme de recherche du projet de remaniement n'avaient pas pour objet d'estimer les économies réalisables sur le plan des coûts. Le personnel chargé des opérations a estimé qu'il y aurait une baisse de 10% dans le coût de la collecte des données, ce qui a permis de calculer la valeur de α .

r : On ne disposait pas des données nécessaires pour estimer ce paramètre, de sorte qu'il a fallu procéder à une simulation de Monte Carlo, décrite à l'annexe B.

3.4 Résultats des analyses du coût et de la variance

Analyse de la variance: Comparaison de P_1 à P_0

Les composantes de la variance de 5 caractéristiques de la population active ont été évaluées pour les plans P_0 et P_1 à partir des données du recensement de 1971 pour 5 RE du Canada. Le tableau 2 indique la proportion de la variance totale attribuable à chaque étape de l'échantillonnage dans le plan P_0 . On peut constater que de 30 à 40% de la variance totale dans P_0 est due aux grappes rurales (3^e étape de l'échantillonnage dans P_0) et que le plan P_1 permettrait une réduction de la variance de l'ordre de 20 à 30%.

i) Stratification : Les parties urbaines et rurales des RE constituerait des strates primaires qui seraient sous-stratifiées optimalement en strates ayant un rendement de 100 à 150 logements (c'est-à-dire en strates composées de 2 ou 3 UPB, dont chacune correspondrait à une tâche d'interviewer). Les RE dans lesquels il n'est pas possible de former ainsi au moins une strate urbaine et une strate rurale (à peu-près 1/3 des RE) seraient considérés comme inadmissibles pour P₂.

Les strates rurales secondaires seraient contiguës, tandis que les strates urbaines secondaires seraient formées sans contraintes géographiques.

ii) Échantillonnage à l'intérieur des strates rurales : Des UPB semblables du point de vue des variables de stratification seraient créées par regroupement de SD géographique-ment contigus, la sélection s'effectuant par la méthode d'ESCAPPT. Les deuxième et troisième étapes de l'échantillonnage seraient la sélection de SD par la méthode d'ESCAPPT et le tirage d'un échantillon systématique de logements.

iii) Échantillonnage à l'intérieur des strates urbaines : L'échantillonnage comprendrait trois étapes : sélection d'UPB (centres urbains séparés ou combinaisons de centres urbains) par la méthode d'ESCAPPT, sélection de grappes par la méthode d'ESCAPPT et choix d'un échantillon systématique de logements.

3.2 Modèle des composantes de la variance

Les plans de sondage P₀, P₁ et P₂ ont été simulés à l'aide des données du recensement. Les formules pour le calcul des composantes de la variance sont les suivantes :

Étape de l'échantillonnage **Formule de variance**

(3.1) $V_{(1)} = V_{ESCAPPT}^{(1)}$ 1ère

(3.2) $V_{(2)} = W_N \sum_{i=1}^I \frac{W_i}{V_{ESCAPPT}^{(2)i}}$ 2e

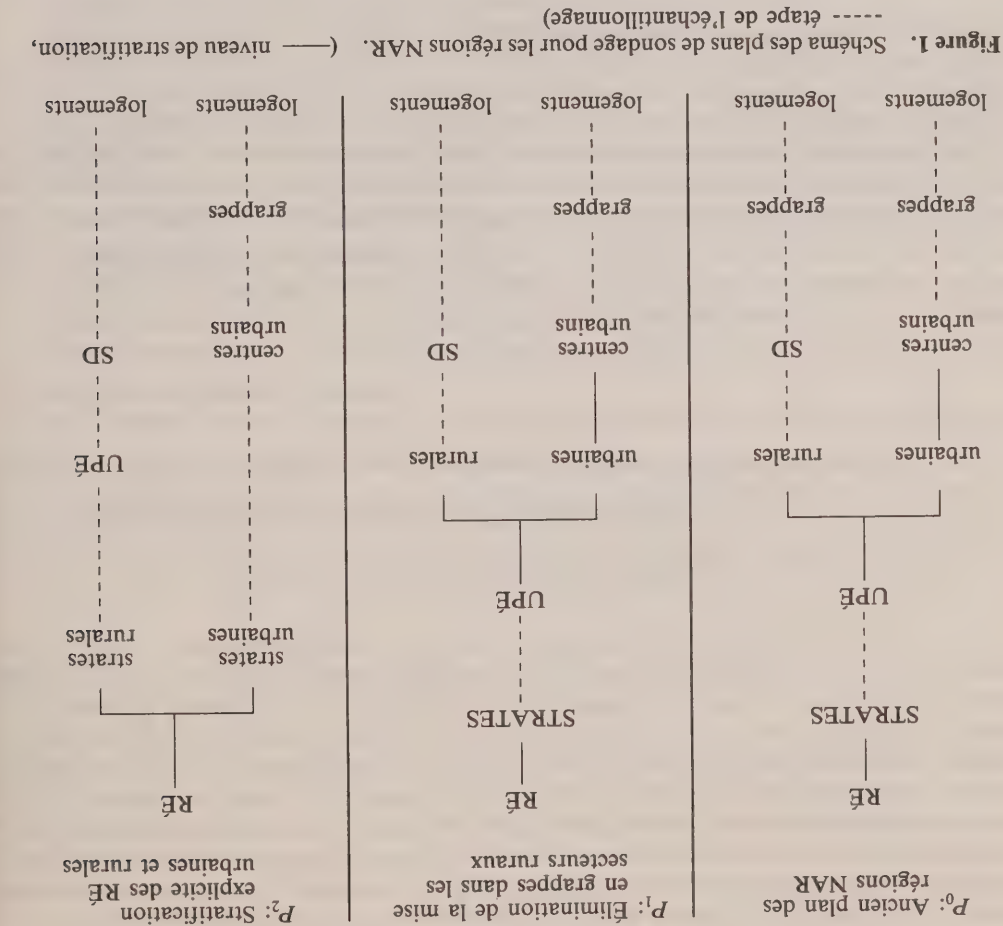
(3.3) $V_{(3)} = W \sum_{j=1}^J \sum_{i=1}^I \frac{W_{ij}}{V_{ESCAPPT}^{(3)ij}}$ 3e
s'il s'agit de la dernière étape
 $= W \sum_{j=1}^J \sum_{i=1}^I \frac{W_{ij}}{V_{ESCAPPT}^{(3)ij}}$ autrement

(3.4) $V_{(4)} = W \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^I \frac{W_{ijk}}{V_{ESCAPPT}^{(4)ijk}}$ 4e (s'il y a lieu)

La formule et la méthode de calcul de la variance pour l'ESCAPPT sont décrites dans l'annexe A.

3.3 Modèle de coût

Le modèle de coût pour les régions AR porte sur la répartition d'échantillons entre les unités de deux degrés du plan de sondage mais, dans le cas des régions NAR, le modèle de coût sert à comparer différents plans de sondage.



- ii) L'élimination de l'étape de la sélection des grappes réduirait les coûts du renouvellement de l'échantillon.
- iii) À priori, un plan d'échantillonnage comportant 3 étapes au lieu de 4 devrait permettre de réduire la variance des estimations. Ce changement ne devrait cependant pas influencer sur les coûts, en particulier à cause de l'utilisation accrue d'interviews téléphoniques.
- iv) Peu de temps après le début du programme de recherche sur le remaniement du plan de sondage de l'EPA, une étude a été menée sur le terrain pour examiner les conséquences de l'élimination de la sélection des grappes sur le déroulement des opérations de collecte des données. Une vérification des listes dressées pour les SD un an après n'a révélé aucun problème dans la qualité de ces listes et une analyse n'a permis de déceler aucun effet sur les coûts de la collecte des données.

Plan de sondage P₂: Stratification explicite des parties urbaines et rurales des UPÉ

Dans l'ancien plan de sondage, l'échantillonnage se fait séparément dans les parties urbaines et rurales des UPÉ, ce qui représente une stratification implicite entre ces parties urbaines et rurales. Un inconvénient de cette méthode est que le maintien à l'échelle des UPÉ du ratio entre la population urbaine et la population rurale pour l'ensemble de la strate entraîne souvent un manque de contiguïté entre les parties rurales et urbaines des UPÉ et, par conséquent, augmente les frais de déplacement.

Comme on pouvait s'y attendre, on observe dans ce modèle que le coût par logement diminue très lentement à mesure que la densité augmente, étant donné que le coût fixe par logement (c_0/m) prédomine dans l'équation (2.10) à cause des interviews téléphoniques. À la section précédente, on a constaté que la variance relative augmente très peu à mesure que la densité augmente. Par conséquent, notre fonction objectif se caractérise par une croissance monotone, mais la perte d'efficacité sur le plan du coût et de la variance qui est provoquée par une augmentation de d est faible. Toutefois, on a décidé de maintenir la valeur de la densité à 5 pour le plan de sondage remanié parce qu'on croit qu'une baisse de la densité entraînerait la sélection d'un grand nombre d'UPÉ pour lesquelles les coûts de la mise sur pied de l'enquête et du renouvellement de l'échantillon seraient très élevés.

3. PLAN DE SONDAGE DES RÉGIONS NAR

3.1 Les différents plans de sondage pour les régions NAR

Plan de sondage P_0 : Ancien plan de sondage des régions NAR (voir figure 1)

Les principales caractéristiques de l'ancien plan de sondage pour les régions NAR (Platek et Singh) sont résumées ci-dessous:

- i) **Stratification:** Les régions économiques (RE), dont le nombre varie de 1 à 10 dans chaque province, servent de grandes strates. À l'intérieur des RE, de 1 à 5 strates géographiquement contiguës sont formées à partir des données sur l'activité industrielle recueillies lors du recensement de 1971.

- ii) **Unités primaires d'échantillonnage (UPÉ):** Les UPÉ ont été définies à l'intérieur de chaque strate sous forme de régions géographiquement compactes semblables à la strate du point de vue des variables de stratification et du ratio entre la population urbaine et la population rurale. Les UPÉ peuvent comprendre de 3,000 à 5,000 habitants. À la première étape d'échantillonnage, on effectue la sélection d'UPÉ par la méthode d'échantillonnage systématique avec classement aléatoire et probabilité proportionnelle à la taille (ESCAPPT) de Hartley et Rao (1962). Dans toutes les UPÉ, les parties urbaines et rurales sont échantillonnées séparément.

- iii) **Echantillonnage à l'intérieur des UPÉ: parties urbaines** Tous les centres urbains dont la totalité ou une partie est affectée aux UPÉ choisies sont inclus dans l'échantillon. La deuxième étape de l'échantillonnage correspond à la sélection d'îlots urbains par la méthode d'ESCAPPT. La troisième et dernière étape de l'échantillonnage comprend le tirage d'un échantillon systématique de logements.

- iv) **Echantillonnage à l'intérieur des UPÉ: parties rurales** La deuxième étape de l'échantillonnage comporte la sélection de secteurs de dénombrement (SD) par la méthode d'ESCAPPT. Le nombre de logements dans les SD est compté sur place afin de définir des grappes composées de 3 à 20 logements. Les troisième et quatrième étapes de l'échantillonnage correspondent au tirage de grappes par la méthode d'ESCAPPT et à la sélection d'un échantillon systématique de logements.

Plan de sondage P_1 : Élimination de l'étape de la sélection des grappes dans les parties rurales des UPÉ

Ce plan de sondage est identique à P_0 , sauf que la sélection des logements se fait directement à l'intérieur des SD ruraux. Il présente des avantages tant sur le plan du déroulement des opérations que sur le plan technique.

- i) Il permettrait de procéder à la sélection d'échantillons indépendants dans la base de sondage de l'EPA 7 mois à l'avance au lieu de 13 mois parce qu'il ne serait plus nécessaire de compter le nombre de logements dans les SD.

Selon les résultats de l'étude des temps et des coûts, les valeurs des paramètres c_1 et c_2 pour Halifax étaient 0.78 et 2.51 respectivement. Ces valeurs ont été enregistrées pour une densité moyenne de 5, mais c_2 augmente quand d augmente alors que c_1 diminue quand d augmente. Supposons que la distance moyenne entre les unités est inversement proportionnelle à la racine carrée du nombre d'unités dans une région. On peut donc remplacer c_1 par $c_1(5/d)^{1/2}$ et c_2 par $c_2(d/5)^{1/2}$ dans notre schéma et on obtient le modèle modifié suivant:

$$(2.10) \quad C = \frac{m}{c_0} + \theta c_1 \left(\frac{d}{5} \right)^{1/2} + \frac{d}{\gamma} \left\{ c_2 \left(\frac{d}{5} \right)^{1/2} - c_1 \left(\frac{d}{5} \right)^{1/2} \right\}.$$

Le rapport c_0/m est le coût fixe par logement et n'est pas lié à la densité; sa valeur était égale à 3.28 selon l'étude des temps et des coûts. Le paramètre θ ne dépend pas non plus de la densité et valait 0.356 selon l'étude des temps et des coûts. Le paramètre γ augmente quand on accroît la densité parce que plus la densité est élevée, plus le nombre moyen de visites à une UPF est élevé. Nous avons utilisé l'approximation suivante de γ :

$$\frac{1}{6} + \frac{5}{2} (1 - d^p)$$

où p est la probabilité que les interviews puissent être menées par téléphone auprès d'un ménage appartenant à une UPF qui n'est à son premier mois d'inclusion dans l'échantillon. La valeur de p était égale à 0.85 selon les données de l'étude sur les interviewers. Les valeurs du coût par logement calculées à partir du modèle (2.10) pour $d = 2, 3, \dots, 10$ sont présentées au tableau 1, où l'on voit également la variance relative des estimations et le produit du coût et de la variance, par rapport auquel la fonction objectif doit être minimisée.

Tableau 1

Valeurs de la variance relative, du coût par logement et de la fonction objectif pour diverses densités (nombre de chômeurs)

Densité	Variance relative	Coût par logement	Fonction objectif
2	0.0206	3.79	0.078
3	0.0214	3.79	0.081
4	0.0222	3.79	0.084
5	0.0230	3.78	0.087
6	0.0238	3.77	0.090
7	0.0246	3.76	0.092
8	0.0254	3.75	0.095
9	0.0262	3.74	0.098
10	0.0270	3.73	0.101

2.2 Modèle de coût

Un modèle de coût simple a été élaboré pour étudier comment les variations dans la densité influent sur le coût. Les interviews dans les régions AR se font par téléphone et les visites sur place sont nécessaires dans une UPF seulement quand un ménage est à son premier mois d'inclusion dans l'échantillon ou dans les cas où un ménage n'a pas de téléphone ou n'a pas consenti à répondre par téléphone.

Les données sur les opérations régionales offrent une ventilation des coûts des interviews faites par téléphone et sur place pour les différents interviews, mais il fallait une analyse plus détaillée des coûts des visites sur place pour qu'on puisse construire le modèle de coût. Pour combler cette lacune, une étude spéciale des temps et des coûts a été entreprise sur le terrain au cours d'une période de six mois (de février à juillet 1982) auprès d'un échantillon aléatoire d'interviewers. Les résultats de l'analyse des données sur les temps et les coûts ont été décrits dans le rapport de Lemaître (1983). Pour notre modèle de coût, nous définissons les paramètres suivants:

c_0 = coût fixe
 c_1 = coût moyen des déplacements d'un logement à l'autre dans une même UPF
 c_2 = coût moyen des déplacements d'une UPF à l'autre
 γ = nombre de déplacements d'une UPF à l'autre par UPF choisie.

Le coût fixe, c_0 comprend le temps que l'interviewer prend pour mener des interviews par téléphone ou sur place et les coûts du déplacement entre le domicile de l'interviewer et le secteur où il travaille et du trajet de retour. Le coût fixe, c_0 dépend seulement de la taille globale de l'échantillon, m , et non de n , le nombre d'UPF choisies. Supposons qu'il y a g_1 déplacements d'un logement à un autre logement et g_2 déplacements d'une UPF à une autre UPF. Le coût total pour m logements est alors

$$T = c_0 + g_1 c_1 + g_2 c_2. \quad (2.7)$$

Si n augmente (diminue), g_2 augmentera (diminuera) aussi et g_1 diminuera (augmentera), mais la somme ($g_1 + g_2$) devrait demeurer constante parce que le nombre de déplacements dépend de la taille de l'échantillon, m et de la proportion de ménages interviewés sur place. On peut donc écrire

$$g_1 + g_2 = \theta m. \quad (2.8)$$

L'équation (2.8) permet de décomposer g_1 dans l'équation (2.7) pour obtenir

$$T = c_0 + \theta m c_1 + g_2 (c_2 - c_1)$$

$$= c_0 + \theta m c_1 + m \gamma (c_2 - c_1).$$

Si on remplace n par m/d on a

$$T = c_0 + \theta m c_1 + \frac{m}{d} (c_2 - c_1)$$

et on peut exprimer le coût par logement, C , en fonction de la densité moyenne:

$$C = \frac{m}{c_0} + \theta c_1 + \frac{d}{\gamma} (c_2 - c_1). \quad (2.9)$$

$$A = \frac{n(N - 1)}{N}$$

La variance relative de Y définie par $\text{Var}(Y)/Y^2$ sera

$$\text{Var rel. } (Y) = A \left[\frac{1}{Y^2} \sum_{j=1}^J \lambda_j^2 - 1 \right] + \frac{1}{Y^2} \sum_{j=1}^J M_j S_j^2 \left[W - 1 - A \left(\frac{\lambda_j}{1} - 1 \right) \right].$$

$$= A\mu_1 + (W - 1)\mu_2 + A\mu_2 - A\mu_3$$

$$= (W - 1)\mu_2 + A(\mu_1 + \mu_2 - \mu_3)$$

(2.5)

où

$$\mu_1 = \frac{1}{N} \sum_{j=1}^J Y_j^2 \lambda_j - 1$$

$$\mu_2 = \frac{1}{N} \sum_{j=1}^J Y_j^2 M_j S_j^2,$$

$$\mu_3 = \frac{1}{S_j^2} \sum_{j=1}^J M_j \lambda_j.$$

μ_1, μ_2 , et μ_3 sont les paramètres de la population et sont fixes pour une caractéristique particulière. Or $m = nd$ et si on suppose que $N_i = N/n$, on peut écrire A de la manière suivante

$$A = \frac{1}{d} \frac{N - 1}{N} \left(\frac{m}{d} - 1 \right)$$

et $\text{Var rel. } (Y) = (W - 1)\mu_2 + (N \frac{d}{m} - 1) \frac{m}{(m_1 + m_2 - m_3)} \frac{(N - 1)}{(N - 1)}$

$$= \alpha_0 + \alpha_1 d$$

(2.6)

où

$$\alpha_0 = (W - 1)\mu_2 - \frac{(N - 1)}{(m_1 + m_2 - m_3)}$$

$$\alpha_1 = \frac{m}{N} \frac{(N - 1)}{(m_1 + m_2 - m_3)}.$$

L'équation (2.6) permet de constater que, du point de vue de la fiabilité des estimations, la valeur $d = 1$ (c'est-à-dire un logement par UPB) est optimale. Toutefois, ce résultat se répercute sur le coût, comme on le verra dans la section suivante. Les valeurs de α_0 et α_1 pour le nombre de chômeurs dans l'UAR de Haïfax ont été calculées à partir des données du recensement de 1981 et on obtient

$$\alpha_0 = 0.019005, \quad \alpha_1 = 0.0007972.$$

Comme α_1 est très faible en comparaison de α_0 , l'accroissement de la variance qui découle d'une augmentation correspondante de d sera très petit. À la section suivante, on examine l'effet des variations dans la valeur de la densité moyenne, d , sur le coût.

$$m = \frac{1}{N} \sum_{j=1}^f M_j = M_0/W \tag{2.2}$$

où M_0 est le nombre total de logements dans la strate. Soit M_j le nombre de logements dans l'UPÉ j choisie dans le $i^{\text{ème}}$ groupe; alors $m_i = M_{ij}/W_{ij}$ logements seront choisis pour le $i^{\text{ème}}$ groupe. Le nombre moyen de logements choisis dans le $i^{\text{ème}}$ groupe pour une répartition donnée des UPÉ en groupes aléatoires est $1/W \sum_j \delta_{ij} M_j$ et la moyenne pour toutes les répartitions possibles des UPÉ en groupes aléatoires est $m N/N$ puisque l'espérance mathématique que de δ_{ij} est N/N . Si $N_i/N = 1/n$, c'est-à-dire, si chaque groupe aléatoire contient le même nombre d'UPÉ, alors l'échantillon moyen par UPÉ choisie sera égal à $m/n = d$, où d représente ce qu'on appellera la densité moyenne pour la strate. Si on fixe m , un échantillon de m logements peut être choisi en variant à la fois n et d de telle façon que le produit (nd) reste égal à m , la taille de l'échantillon tiré de la strate. Notre objectif ici est d'obtenir la valeur de d qui, pour un échantillon de taille fixe, permet de minimiser le produit de la variance et du coût. Pour obtenir la solution optimale, nous calculons la variance totale en procédant à une analyse des composantes de la variance et nous examinons une fonction de coût linéaire. Ces aspects sont décrits dans la section suivante.

2.1 Fonction de variance

Supposons que nous voulons estimer la valeur totale d'une caractéristique y dans une sous-unité. Soit y_{jh} la valeur de y pour le $h^{\text{ème}}$ ménage de l'UPÉ j où $h = 1, 2, \dots, N$, on peut estimer le total $Y = \sum_{j=1}^f \sum_{h=1}^{M_j} y_{jh}$ à l'aide de la variable

$$Y = W \sum_{i=1}^n y_i \tag{2.3}$$

où y_i est la somme des valeurs de y pour les m_i ménages choisis dans l'UPÉ tirée du $i^{\text{ème}}$ groupe, $i = 1, 2, \dots, n$. Abstraction faite de l'effet dû à l'arrondissement dans la définition de W_{ij} , la formule de la variance de Y est (Rao et coll., 1962)

$$\text{Var}(Y) = A \left[\sum_{j=1}^f \frac{\lambda_j}{Y_j^2} - Y^2 \right] + \sum_{j=1}^f M_j S_j^2 \left[W - 1 - A \left(\frac{1}{Y_j} - 1 \right) \right]. \tag{2.4}$$

où

$$Y_j = \sum_{h=1}^{M_j} y_{jh},$$

$$S_j^2 = \frac{1}{M_j} \sum_{h=1}^{M_j} \left(y_{jh} - \frac{Y_j}{M_j} \right)^2,$$

$$A = \frac{\sum_{i=1}^n N_i^2 - N^2}{N(N-1)}.$$

Si $N_i = N/n$, c'est-à-dire si tous les groupes aléatoires ont le même nombre d'UPÉ, alors

Cette analyse vise à choisir la répartition optimale de l'échantillon entre les deux degrés du plan de sondage à l'intérieur des régions AR (section 2) et à évaluer deux plans de sondage susceptibles de remplacer l'ancien plan à l'intérieur des régions NAR (section 3). Pour les régions NAR, on évalue d'abord l'ancien plan de sondage empiriquement par une analyse des composantes de la variance et on désigne une étape de l'échantillonnage à éliminer dans les régions rurales. Ensuite, du point de vue général du coût et de la variance, on compare l'ancien plan de sondage remanié avec un autre plan de sondage comportant une stratification explicite des régions urbaines et rurales. Des variances sont estimées empiriquement pour les régions AR et NAR à partir des données des recensements de 1971 et 1976 et des modèles de coût sont construits à l'aide des données d'une étude des temps et des coûts et des résultats d'une étude de simulation.

Dans la section 4, nous examinons la deuxième étape de l'optimisation, la répartition de l'échantillon entre les régions AR et NR, tout en tenant compte des améliorations apportées au plan de sondage dans ces deux types de région. Enfin, la section 5 résume les améliorations décrites dans cet article et leurs répercussions sur le remaniement de l'échantillon.

2. PLAN DE SONDAGE DES RÉGIONS AR

L'ancien échantillon des régions AR repose sur un plan de sondage stratifié à deux degrés (Platek et Singh, 1976). Chaque région AR est stratifiée en un certain nombre de strates constituées de "sous-unités" et chaque sous-unité est divisée en grappes qui constituent les unités primaires d'échantillonnage (UPF). Les UPF sont choisies à l'aide de la méthode des groupes aléatoires élaborée par Rao, Hartley et Cochran (1962) et, à la deuxième étape de l'échantillonnage, un échantillon systématique de logements est pris de manière que le plan de sondage soit autopondéré. Soit $1/W$ la fraction de sondage dans une strate donnée et soit n le nombre d'UPF qui doivent être choisies dans cette strate. Les N UPF de la strate sont réparties au hasard en n groupes de telle sorte que le $j^{\text{ème}}$ groupe aléatoire contient N_j UPF et $\sum_{j=1}^n N_j = N$. Soient x_j et M_j , $j = 1, 2, \dots, N$, respectivement la taille de la $j^{\text{ème}}$ UPF dans la strate et le nombre de logements dans cette UPF.

Définitions

$$\lambda_j = \frac{\sum_{i=1}^I x_i}{x_j}$$

et $\delta_{ij} = 1$ si la $j^{\text{ème}}$ UPF est dans le $i^{\text{ème}}$ groupe, $= 0$ autrement.

Alors $\pi_i = \sum_{j=1}^n \delta_{ij} \lambda_j$ est la taille relative du $i^{\text{ème}}$ groupe. Définissons des W_j de la manière suivante:

$$W_j = \delta_{ij} \left[W \frac{\pi_i}{\lambda_j} \right] \text{ ou } \delta_{ij} \left[W \frac{\pi_i}{\lambda_j} + 1 \right]$$

de sorte que $\sum_{j=1}^n W_j = W$ pour $i = 1, 2, \dots, n$, où $[a]$ est l'entier le plus élevé qui est inférieur ou égal à a . Choisissons une UPF indépendamment dans chacun des n groupes aléatoires avec une probabilité proportionnelle aux W_j et choisissons ensuite un sous-échantillon dans la $j^{\text{ème}}$ UPF du $i^{\text{ème}}$ groupe aléatoire avec fraction de sondage égale à $1/W_j$. La fraction de sondage globale dans chaque groupe aléatoire est alors égale à $1/W$ et l'échantillon est donc autopondéré avec un poids égal à W . La taille moyenne des échantillons choisis dans la strate est

Optimisation du coût et de la variance dans le cadre de l'enquête sur la population active au Canada

G.H. CHOUDHRY, H. LEE, et J.D. DREW¹

RÉSUMÉ

L'optimisation du coût et de la variance dans le remaniement du plan de sondage de l'enquête sur la population active au Canada s'est déroulée en deux étapes. À la première étape, le plan de sondage a été optimisé par rapport à chacune des deux grandes catégories d'unités géographiques, les régions autorenseignables (AR) et les régions non autorenseignables (NAR). On a construit des modèles de coût, estimé leurs paramètres à partir des résultats d'une étude détaillée menée sur le terrain et de simulations et calculé des variances à l'aide des données du recensement de la population. Cette analyse visait également à choisir la répartition optimale de l'échantillon entre les deux degrés du plan de sondage à l'intérieur des régions AR et à évaluer deux plans de sondage susceptibles de remplacer l'ancien plan à l'intérieur des régions NAR. À la deuxième étape, la répartition optimale de l'échantillon entre les régions AR et NAR a été déterminée.

MOTS CLÉS: Plans de sondage à plusieurs degrés; répartition de l'échantillon; fonction de coût linéaire; composantes de la variance.

1. INTRODUCTION

Statistique Canada mène l'enquête sur la population active au Canada (EPA) tous les mois auprès d'un échantillon de ménages pour produire des estimations de diverses caractéristiques de la population active. Cette enquête est fondée sur un plan de sondage à plusieurs degrés et un plan de renouvellement en vertu duquel l'échantillon est divisé en six groupes. Depuis le début de l'EPA en 1945, le plan de sondage est remanié après chaque recensement décennal de la population. Cette opération permet de mettre à jour l'échantillon de manière à tenir compte des changements démographiques. Elle permet également d'introduire des techniques améliorées d'échantillonnage et d'estimation et de suivre l'évolution des besoins en information auxquels l'EPA doit répondre.

Les travaux de remaniement entrepris après le recensement de 1981 comprenaient une phase de recherche qui a été résumée par Singh et Drew (1981) et au cours de laquelle tous les aspects du plan de sondage ont été examinés dans le but d'améliorer le rapport entre le rendement et le coût de l'EPA. Les faits saillants de ce programme de recherche ont été décrits par Singh, Drew et Choudhry (1984). Le présent article porte sur les analyses concernant l'optimisation du coût et de la variance dans le plan de sondage.

Les deux principaux facteurs qui déterminent le choix d'un plan de sondage sont le coût total et la fiabilité des estimations calculées à partir de l'échantillon. Pour obtenir la solution optimale, on peut minimiser soit le coût total pour une variance fixe, soit la variance totale pour un coût fixe. Nous avons adopté une méthode équivalente dans laquelle on minimise le produit de la variance et du coût pour un échantillon d'une taille donnée. L'optimisation du coût et de la variance s'est déroulée en deux étapes. Nous examinons d'abord l'optimisation des plans de sondage utilisés pour chacune des deux grandes catégories d'unités géographiques définies pour l'EPA, c'est-à-dire les régions AR, ou les grandes villes, et les régions NAR, qui correspondent aux petites régions urbaines et aux régions rurales.

¹ G.H. Choudhry, H. Lee, et J.D. Drew, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, 4ième étage, Édifice Jean Talon, Parc Tunney, Ottawa, Ontario K1A 0T6.

- DOSS, D.C., HARTLEY, H.O., et SOMAYAJULU, G.R. (1979). An exact small sample theory for post-stratification. *Journal of Statistical Planning and Inference*, 3, 235-248.
- DREW, J.H., SINGH, M.P., et CHOUDHRY, H. (1982). Evaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active au Canada. *Techniques d'enquêtes*, 8, 17-47.
- DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. Dans *New Developments in Survey Sampling* (Eds. N.L. Johnson et H. Smith), New York: Wiley - Interscience.
- FISHER, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd (5e éd., 1934).
- FULLER, W.A. (1966). Estimation employing post strata. *Journal of the American Statistical Association*, 61, 1172-1183.
- HARTLEY, H.O., RAO, J.N.K., et KIEFFER, G. (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association*, 64, 841-851.
- HIDIROGLOU, M.H., et SÄRNÄDAL, C.E. (1985). Etude empirique de quelques estimateurs de régression pour petits domaines. *Techniques d'enquête*, 11, 65-77.
- HIDIROGLOU, M.H., et SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.
- HOLT, D., et SMITH, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Série A*, 142, 33-46.
- LAHIRI, D.B. (1969). On the unique sample, the surveyed one. Document technique non publié, Indian Statistical Institute.
- OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys*, vol. 2, Academic Press, 142-184.
- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- SÄRNÄDAL, C.E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- WILLIAMS, W.H. (1962). The variance of an estimator with post-stratified weighting. *Journal of the American Statistical Association*, 57, 622-627.
- YATES, F. (1984). Tests of significance for 2×2 contingency tables. *Journal of the Royal Statistical Society, Série A*, 147, 426-463.

sinon $y_i(i) = 0$, pourrait mieux convenir que $v^*(Y^{pst})$ puisque, dans le cas d'un échantillon-
nage aléatoire simple il se réduit à (3.3) lorsque multiplié par N^2 et qu'il n'est pas néces-
saire d'introduire une correction d'échantillonnage pour population finie

$$v(Y^{pst}) = \sum_i^I \frac{N_i^2}{s_{iy}^2} \tag{7.10}$$

En outre, d'après une théorie des estimateurs par quotient appliqués à des modèles, $v(Y^{pst})$
serait conditionnellement plus efficace que $v^*(Y^{pst})$. Quoiqu'il en soit, il n'y a pas de mal
à utiliser l'estimateur (7.8) puisque, en l'absence de toute condition, il est asymptotiquement
équivalent à l'estimateur habituel de la variance (équation 7.5).

8. ANALYSE

La présente étude montre clairement que l'inférence conditionnelle dans les plans de son-
dage complexes soulève d'énormes problèmes. Il ne faut pas pour autant utiliser aveuglè-
ment les méthodes classiques. Lorsque l'inférence conditionnelle est possible, comme dans
le cas de l'échantillonnage aléatoire simple, nous devons certes utiliser des méthodes adap-
tées à un cadre d'analyse conditionnel, comme celles qui sont décrites dans les sections 2
à 6, tandis que dans les cas plus complexes, nous devons au moins modifier légèrement
les méthodes traditionnelles, comme dans le cas de l'équation (7.6), afin qu'elles soient com-
patibles avec les résultats conditionnellement justes observés dans des cas particuliers. De
toute évidence, ce domaine doit faire l'objet d'autres recherches.

REMERCIEMENTS

Ce document est basé sur les exposés que j'ai faits à un atelier de travail sur l'inférence
conditionnelle dans les enquêtes par sondage. Je tiens à remercier Mme Nanjamma Chin-
nappa pour avoir organisé l'atelier à Statistique Canada. Je remercie également mes collè-
gues de Statistique Canada et le professeur D. Holt pour les commentaires utiles et construc-
tifs qu'ils m'ont donnés au cours de la rédaction de ce document.

BIBLIOGRAPHIE

BANKIER, M. (1985). Conditionally unbiased estimators based on any number of independent strati-
fied samples. Mémoire, Division des méthodes d'enquêtes-entreprises, Statistique Canada.
BATTSE, G.E., et FULLER, W.A. (1981). Prediction of county crop areas using survey and satel-
lite data. *Proceedings of the Section on Survey Research Methods, American Statistical Associa-
tion*, 500-505.
BRYANT, E.C., HARTLEY, H.O., et JESSEN, R.J. (1960). Design and estimation in two-way stra-
tification. *Journal of the American Statistical Association*, 55, 105-124.
CHINNAPPA, B.N. (1976). A preliminary note on methods of dealing with unusually large units in
sampling from skew populations. Document technique non publié, Division des méthodes d'enquêtes-
institutions et agriculture, Statistics Canada.
COX, D.R., et HINKLEY, D.V. (1974). *Theoretical Statistics*. Londres, Chapman et Hall.
DEMPSTER, A.P., RUBIN, D.B., et TSUTAKAWA, R.K. (1981). Estimation in covariance compo-
nent models. *Journal of the American Statistical Association*, 76, 341-353.

où $N_{hi} = N_{h1} + N_{h2}$ et $n_{hi} = n_{h1} + n_{h2}$ désignent respectivement la taille de la population et la taille de l'échantillon de la strate et où y_{hi} est le total de l'échantillon dans la (h, i) -ième cellule. Étant donné (n_1, n_2) , l'espérance conditionnelle de (7.2) ne peut être déterminée puisqu'il faut faire la somme

$$(7.3) \quad E_2(Y^{pst}) = \sum_i p(s_i | n_1, n_2) Y^{pst}_i(t)$$

où s_i est un échantillon probable tel que la taille observée n_{hi} des échantillons satisfait l'équation $n_{h1} + n_{h2} = n_i$ ($i = 1, 2$), et $Y^{pst}_i(t)$ est la valeur de l'équation (7.2) pour l'échantillon s_i , et $p(s_i | n_1, n_2)$ est la probabilité conditionnelle que s_i soit observé étant donné (n_1, n_2) :

$$(7.4) \quad p(s_i | n_1, n_2) = \left[\sum_{n_{11}=0}^{n_{11}} \begin{pmatrix} N_{11} \\ n_{11} \end{pmatrix} \begin{pmatrix} N_{12} \\ n_{11} - n_{11} \end{pmatrix} \begin{pmatrix} N_{21} \\ n_{11} - n_{11} \end{pmatrix} \begin{pmatrix} N_{22} \\ n_2 - n_{11} + n_{11} \end{pmatrix} \right]^{-1}.$$

Cependant, les équations (7.3) et (7.4) nous indiquent clairement que $E_2(Y^{pst}) \neq Y$ puisque, contrairement à $p(s_i | n_1, n_2)$, Y^{pst} ne dépend pas des totaux N_{hi} des cellules. Pour ce qui a trait à l'estimation de la variance, la formule utilisée habituellement dans les plans de sondage généraux est exprimée:

$$(7.5) \quad v^*(Y^{pst}) = v(z_i^*)$$

où $v(y_i) = v(Y)$ est l'estimateur habituel de la variance du total estimé Y ; on obtient $v(z_i^*)$ à partir de $v(Y)$ en remplaçant y_i par

$$(7.6) \quad z_i^* = y_i - \sum_i \frac{M_i}{Y} a_i(t)$$

où $a_i(t) = 1$ si le i -ième élément appartient à la i -ième strate formée a posteriori; sinon $a_i(t) = 0$ (Williams 1962). Dans le cas d'un échantillonnage aléatoire simple, l'équation (7.5) devient

$$(7.7) \quad v^*(Y^{pst}) = N^2 \left(\frac{1}{N} - \frac{1}{N^2} \right) \sum n_{st}^2$$

(en supposant que $(n_1 - 1)/(n - 1) = n_1/n$); l'équation ci-dessus n'est pas égale à l'équation (3.3) lorsqu'elle est multipliée par N^2 . Donc, l'équation (7.5) ne convient pas très bien au cadre d'analyse conditionnel, même lorsqu'il s'agit d'un échantillonnage aléatoire simple. par ailleurs, un nouvel estimateur de variance

$$(7.8) \quad v(Y^{pst}) = v(z_i),$$

où

$$(7.9) \quad z_i = \sum_i \frac{M_i}{Y} (y_i(t) - \frac{M_i}{Y} a_i(t))$$

et $y_i(t) = y_i$ si le i -ième élément appartient à la i -ième strate formée a posteriori,

À l'aide des renseignements préables, il est possible de modifier Y_i :

$$(6.11) \quad Y_i^* = \begin{cases} N_i^* y_i & \text{si } N_i < N_i^* \\ N_i y_i & \text{si } N_i^* \leq N_i \leq N_i^{**} \\ N_i^{**} y_i & \text{si } N_i > N_i^{**} \end{cases}$$

Le biais conditionnel de Y_i^* est inférieur, en valeur absolue, à celui de Y_i si $N_i < N_i^*$ ou $N_i > N_i^{**}$, tandis que $Y_i^* = Y_i$ dans l'intervalle $N_i^* \leq N_i \leq N_i^{**}$. En conséquence, Y_i^* est conditionnellement plus efficace que l'estimateur non biaisé Y_i . De plus, l'EQM inconditionnelle de Y_i^* est moindre que celle de Y_i , même si Y_i^* est inconditionnellement biaisé. Malheureusement, il n'existe pas de moyen facile d'améliorer l'efficacité de Y_i^* dans l'intervalle $N_i^* \leq N_i \leq N_i^{**}$. Quoi qu'il en soit, Y_i^* devrait être préféré à Y_i pour que l'estimation du total d'un domaine soit efficiente, il faut disposer de bons renseignements supplémentaires sur la taille du domaine.

7. PLANS DE SONDAGE GÉNÉRAUX

L'ajustement par stratification a posteriori est couramment utilisé dans les grandes enquêtes complexes afin, surtout, d'accroître l'efficacité des estimateurs. Mentionnons, à titre d'exemple, le facteur âge-sexe dans l'enquête sur la population active du Canada (EPA). Il existe également une théorie générale de l'inférence inconditionnelle. L'estimateur du total Y est défini par :

$$(7.1) \quad Y_{pst} = \sum M_i \frac{\bar{M}_i}{Y_i}$$

où Y_i et M_i sont respectivement les estimateurs non biaisés habituels du total Y_i et de la taille M_i de la i -ième strate formée a posteriori. Dans l'EPA, on se sert des données projetées du recensement pour les M_i . L'estimateur Y_{pst} se réduit à $\sum N_i y_i$ dans le cas d'un échantillonnage aléatoire simple (voir (3.2)), et nous avons vu précédemment que $\sum N_i y_i$ est conditionnellement non biaisé dans un échantillon aléatoire simple (en supposant que $n_i \geq 1$ pour tous les i). Toutefois, dans les plans de sondage complexes, il semble difficile d'analyser les propriétés conditionnelles de (7.1); même le choix d'un ensemble de référence n'est pas aussi évident que l'on croit. Pour illustrer ce problème, considérons un échantillon-nage aléatoire simple stratifié où $L = 2$ strates et $k = 2$ strates formées a posteriori. Si nous faisons reposer l'inférence conditionnelle sur les tailles d'échantillon (n_{h1}, n_{h2}) observées dans chaque strate h formée a posteriori, la théorie s'applique normalement pourvu que l'on connaisse les tailles N_{hi} des strates formées a posteriori. En pratique, toutefois, nous risquons d'éprouver des difficultés si nous avons des tailles n_{hi} d'échantillon nulles; il se peut aussi que nous ne connaissions pas les tailles N_{hi} des strates ou que les projections soient inexactes même si nous avons $N_i = \sum_h N_{hi} = M_i$. Il serait alors préférable de faire reposer l'inférence conditionnelle sur la taille totale observée (n_1, n_2) des échantillons, où $n_i = \sum_h n_{hi}$.

Dans ce cas particulier d'échantillonnage aléatoire simple stratifié ($L = 2, k = 2$), l'estimateur Y_{pst} s'exprime comme suit :

$$(7.2) \quad Y_{pst} = N_1 \frac{Y_{11} \frac{N_1 n_1}{n_{11}} + N_2 \frac{N_1 n_1}{n_{21}}}{N_1 \frac{n_1}{n_{11}} + N_2 \frac{n_1}{n_{21}}} + N_2 \frac{Y_{12} \frac{N_1 n_1}{n_{12}} + N_2 \frac{N_1 n_1}{n_{22}}}{N_1 \frac{n_1}{n_{12}} + N_2 \frac{n_1}{n_{22}}}$$

Drew et al. (1982) ont proposé un autre estimateur qui dépend de la taille de l'échantillon ainsi que d'un paramètre K_0 . Dans le cas de l'échantillonnage aléatoire simple, si l'on choisit $K_0 = 1$, l'estimateur en question devient:

$$(6.7) \quad y_{id} = \begin{cases} y_i & \text{si } w_i \geq W_i \\ y_{is} & \text{si } w_i < W_i \end{cases}$$

Nous avons souligné plus haut que le choix entre y_{is} et y_i^* ne s'imposait pas à l'évidence lorsque $w_i < W_i$. Il en va de même du choix entre y_{id} et y_{is}^* .

Si N_i est inconnue, l'inférence conditionnelle est encore possible pourvu que N_i n'ait aucun rapport avec le paramètre étudié X_i . Elle est également faisable lorsqu'on dispose de renseignements partiels sur N_i (si l'on connaît, par exemple, les bornes de N_i).

S'il existe une variable concomitante x ayant une moyenne de domaine X_i connue, l'estimateur par quotient

$$(6.8) \quad y_{ir} = \frac{x_i}{X_i}$$

et l'estimateur par régression (Battese et Fuller 1981)

$$(6.9) \quad y_{ir}^* = y_i + \frac{x}{X} (X_i - x_i)$$

sont tous deux conditionnellement non biaisés (approximativement), mais y_{ir}^* sera vraisemblablement plus efficace s'il est possible d'appliquer, du moins approximativement, une équation de régression de même pente (dont la courbe passe par l'origine) aux petites régions. Si les pentes diffèrent, il conviendrait mieux d'utiliser un estimateur empirique de Bayes, qui est plus complexe (Dempster et coll., 1981).

6.2 Total d'un domaine

Si N_i est connue, on peut facilement obtenir une estimation du total d'un domaine $Y_i = N_i Y_i$ en multipliant un estimateur choisi de Y_i par N_i . Si, par ailleurs, N_i est inconnue, on se sert de l'estimateur non biaisé habituel

$$(6.9) \quad Y_i = N_i y_i = \frac{n}{N} \sum_{j \in s_i} Y_j, \quad n_i \geq 1$$

où $N_i = N w_i$ est l'estimateur sans biais de N_i et $P(n_i = 0)$ est supposé négligeable.

Supposons, maintenant, que nous disposions de renseignements préalables, par exemple $N_i^* \leq N_i \leq N_i^{**}$; l'inférence conditionnelle est alors possible. Le biais conditionnel de Y_i est

$$(6.10) \quad B_2(Y_i) = (N_i^* - N_i) Y_i$$

L'équation ci-dessus nous permet d'affirmer (en supposant que $Y_i > 0$) que $B_2(Y_i) > 0$, c'est-à-dire, qu'il y a surestimation si $N_i^* > N_i$ et que $B_2(Y_i) < 0$, c'est-à-dire qu'il y a sous-estimation si $N_i^* < N_i$ et $|B_2(Y_i)|$ augmente lorsque n_i diminue; le biais conditionnel est nul si $N_i = N_i^*$.

où $a_i = 1$ si $n_i \geq 1$; $= 0$ si $n_i = 0$ et où y_i est assimilée à X_i si $n_i = 0$. L'estimateur défini en (6.3) est toutefois conditionnellement biaisé:

$$E_2(y_i') = \frac{E(a_i)}{a_i} Y_i$$

Il constitue une sous-estimation si $n_i = 0$, et une surestimation si $n_i \geq 0$, même s'il est inconditionnellement non biaisé. Le degré de surestimation dépend de la valeur de $E(a_i) = P(n_i \geq 1)$. Si, par exemple, $P(n_i \geq 1) = 0.75$, alors $E_2(y_i') = (\frac{4}{3}) Y_i$ si $n_i \geq 1$. Sarnadal (1984) a proposé l'estimateur suivant dans le cas de l'estimation pour les petites régions:

$$y_{is} = y + \frac{W_i}{w_i} (y_i - y), n_i \geq 0, \tag{6.4}$$

où $y = \sum w_j y_j$ est la moyenne globale de l'échantillon et $w_i = n_i/n$. L'estimateur est à peu près inconditionnellement non biaisé, mais il est conditionnellement biaisé à moins que

$$B_2(y_{is}) = \left(\frac{W_i}{w_i} \right) - 1 (Y_i - Y'), \tag{6.5}$$

où $Y' = \sum w_i X_i$. Si $n_i = 0$, l'estimateur y_{is} se réduit à l'estimateur synthétique y . Le degré de sous-estimation (ou de surestimation) de y_{is} dépend à la fois de $w_i/W_i - 1$ et de $Y_i - Y'$ et est, par conséquent, plus difficile à analyser que le biais de y_i' . Cependant, le biais conditionnel* de y_{is} serait supérieur, en valeur absolue, à celui de y si $w_i > 2W_i$ (l'EQM conditionnelle de y_{is} serait de ce fait plus grande). De plus, la variance conditionnelle de l'estimateur conditionnellement non biaisé y_i' est moindre que celle de y_{is} si $w_i > W_i$ (si l'on ne tient pas compte de la variance de y par rapport à celle de y_i) et, de ce fait, l'EQM conditionnelle de y_i' est moindre que celle de y_{is} .

Hidiroglou et Sarnadal (1985) proposent une variante de y_{is} :

$$y_{is}^{**} = \begin{cases} y_i \text{ si } w_i \geq W_i \\ y_{is}^* = y + \left(\frac{W_i}{w_i} \right)^2 (y_i - y) \text{ si } w_i < W_i. \end{cases} \tag{6.6}$$

L'estimateur y_{is}^{**} est conditionnellement non biaisé si $w_i \geq W_i$, tandis que son biais conditionnel, en valeur absolue, est moindre que celui de y si $w_i < W_i$. L'estimateur défini en (6.6) est justifié par le fait que la variance conditionnelle de y_{is}^* (ou y_{is}) est supérieure à celle de y_i' (si l'on ne tient pas compte de la variance de y par rapport à celle de y_i), si $w_i > W_i$, tandis qu'elle est inférieure à celle de y_{is} si $w_i < W_i$. Sous cette condition, toutefois, le biais conditionnel de y_{is}^* est supérieur, en valeur absolue, à celui de y_{is} . Ainsi, le choix entre ces deux estimateurs lorsque $w_i < W_i$ ne s'impose pas à l'évidence, et il ne semble y avoir de solution simple.

* L'estimateur de Sarnadal devrait toutefois être plus efficace dans le cas d'une stratification simple. On obtient cet estimateur en groupant des estimateurs comme celui défini en (6.4) qui s'appliquent à deux groupes ou plus.

déterminé m_i (Oh et Scheuren 1983). En conséquence, l'estimateur

(5.4)
$$y^{pst,m}_i = \sum W_j y^{mj}_i$$

est conditionnellement non biaisé et sa variance conditionnelle est estimée sans distorsion par l'équation suivante;

(5.5)
$$v_2(y^{pst,m}_i) = \sum W_j^2 \left(\frac{1}{s^2_{mj}} m_i - \frac{1}{s^2_{mj}} m_i \right)$$

où y^{mj}_i et s^2_{mj} représentent respectivement la moyenne et la variance de l'échantillon des répondants dans la ième strate formée a posteriori.

Si on ne connaît pas les valeurs de W_j , il est de pratique courante de remplacer W_j dans l'équation (5.4) par son estimation $w_j = n_j/n$. On peut alors s'interroger sur la valeur de l'inférence conditionnelle puisque la distribution de (n_j, m_j) dépend des poids W_j , que l'on ne connaît pas, et que ces poids entrent dans le calcul du paramètre $\bar{Y} = \sum W_j \bar{Y}_j$. Si l'on disposait de données partielles sur W_j (par exemple, les bornes de W_j), l'inférence conditionnelle pourrait se faire comme dans le cas décrit à la remarque 3 de l'exemple 1, cependant on obtiendrait encore un estimateur conditionnellement biaisé (mais vraisemblablement plus efficace que l'estimateur (5.4) où W_j est remplacé par w_j).

6. ESTIMATION DE DOMAINES (EAS)

6.1 Moyenne d'un domaine

Dans un échantillonnage aléatoire simple (EAS), l'estimateur habituel de la moyenne d'une sous-population (domaine), \bar{Y}_i , est exprimé par la moyenne de l'échantillon

(6.1)
$$y_i = \sum_{j \in s_i} \frac{n_j}{n_i}, n_i > 0$$

où s_i est l'échantillon prélevé dans le domaine et n_i est la taille correspondante.

Si la taille du domaine, N_i , est connue, on devrait faire reposer l'inférence sur la valeur observée n_i . L'estimateur y_i est conditionnellement non biaisé si $n_i > 0$ puisque, conditionnellement, s_i est un échantillon aléatoire simple de taille déterminée n_i prélevé dans le domaine. La variance conditionnelle de cet estimateur est estimée sans distorsion par:

(6.2)
$$v(y_i) = \frac{1}{n_i} - \frac{1}{N_i} \left(s^2_{iy}, n_i > 0 \right)$$

et l'intervalle de confiance correspondant $y_i \pm z_{\alpha/2} \sqrt{v(y_i)}$ est conditionnellement juste.

L'estimateur y_i est toutefois irrégulier dans le cas de petits domaines (petites régions) où n_i est petit. De plus, y_i n'est pas défini si $n_i = 0$. Pour remédier à cela, les statisticiens proposent d'utiliser un estimateur modifié

(6.3)
$$y'_i = \frac{E(a_i)}{a_i} y_i, n_i \geq 0$$

répondants est inconditionnellement biaisée puisque $E(Y^m) = Y_2 \neq Y$. Il est donc nécessaire de construire un modèle de réponse, même en l'absence de toute condition, à moins qu'on ne prélève un sous-échantillon de non-répondants. Dans un modèle simplifié, on suppose que la probabilité d'une réponse est la même pour toutes les unités rejoindtes, disons p^* ; en d'autres termes, les données manquantes sont réparties aléatoirement. Suivant ce modèle, la distribution de m dépend uniquement de p^* et nous devrions, par conséquent, faire reposer l'inférence sur m si p^* est supposée connue (ou, du moins, partiellement connue ou non liée à Y). Oh et Scheuren (1983) ont montré que, moyennant une valeur m donnée, l'échantillon s_m de répondants pouvait être comparé à un échantillon aléatoire simple de taille m prélevé dans la population globale. Ainsi, y^m est conditionnellement non biaisé et sa variance conditionnelle est estimée sans distorsion par l'équation suivante:

(5.1)
$$v_2(y^m) = (m^{-1} - N^{-1})s_{my}^2$$

où $(m - 1)s_{my}^2 = \sum_{i \in s_m} (y_i^m - \bar{y}^m)^2$. L'intervalle de confiance correspondant $\bar{y}^m \pm z_{\alpha/2} \sqrt{v_2(y^m)}$ est conditionnellement juste, du moins approximativement, si m est grand.

Par ailleurs, l'estimateur d'Horvitz-Thompson (p^* étant connue):

(5.2)
$$y_{HT} = \frac{E(m)}{m} \bar{y}^m = \frac{\sum_{i \in s_m} m p^*}{y_i}$$

est conditionnellement biaisé, comme dans le cas illustré à la section 2, mais il est non biaisé lorsqu'il s'applique à l'ensemble de la distribution de m . En ce qui concerne les plans de sondage généraux, l'estimateur par quotient

(5.3)
$$\hat{Y}_{HT,r} = \frac{\sum_{i \in s_m} \frac{\pi_i d_i^*}{y_i}}{\sum_{i \in s_m} \frac{\pi_i d_i^*}{1}} = \frac{\sum_{i \in s_m} \pi_i d_i^*}{1}$$

est souvent utilisé pour des raisons d'efficacité; dans l'équation ci-dessus, π_i est la probabilité d'inclusion et p^* est la probabilité de réponse (supposée connue) de la i -ième unité à condition que celle-ci ait été contactée. Dans le cas d'un échantillonnage aléatoire simple où $p^* = p$, il est intéressant de constater que $\hat{Y}_{HT,r}$ se réduit à \bar{y}^m . On peut donc en déduire que l'estimateur par quotient pourrait s'avérer efficace dans un cadre d'analyse conditionnel appliqué à des plans de sondage généraux.

5.2 Modèle plus conforme à la réalité

Dans un modèle plus conforme à la réalité, on suppose que les données manquantes sont réparties aléatoirement dans des strates formées à posteriori et pondérées par des poids W_i connus. Soit n_i et m_i qui désignent respectivement la taille de l'échantillon et la taille de l'échantillon des répondants dans la i -ième strate formée à posteriori. Alors, la distribution à plusieurs variables de (n_i, m_i) dépend uniquement de la valeur de W_i et des probabilités de réponse dans les strates formées à posteriori. L'inférence devrait donc reposer sur les valeurs observées de (n_i, m_i) , pourvu que les probabilités de réponse dans les strates formées à posteriori soient connues ou qu'elles n'aient aucun rapport avec les paramètres étudiés (par exemple, les moyennes des strates formées à posteriori). Conditionnellement, l'échantillon observé est comparable à un échantillon aléatoire simple stratifié avec des strates de taille

et

$$G_j(p) = \frac{\sum_{i=1}^j n_i}{n} = W_j = \sum_{i=1}^j W_{ji}$$

Les valeurs de $G_j(p)$ sont déterminées comme suit: soit $G_j(0) = G_j > 0 \forall (i, j)$, et

$$G_j(p) = G_j(p - 1) \frac{\sum_{i=1}^j \frac{n_i}{G_j(p - 1)}}{W_j} \quad \text{si } p \text{ est impair} \quad (4.6)$$

$$= G_j(p - 1) \frac{\sum_{i=1}^j \frac{n_i}{G_j(p - 1)}}{W_j} \quad \text{si } p \text{ est pair.}$$

En vertu du modèle des effets permanents (4.3), nous avons

$$E_m[Y(p)] = \mu + \sum_{i=1}^j W_{ji} \tau_i + \sum_{j=1}^j W_{jj} \beta_j = E_m(\sum_{j=1}^j W_{jj} Y_{jj})$$

$$= E_m(Y),$$

c'est-à-dire que $y(p)$ n'est à peu près pas biaisé en fonction du modèle. Puisque $E(G_j(0)n_j/n) = W_j$ pour la sélection $G_j(0) = G_j$, les valeurs initiales devraient être bonnes. La méthode itérative pourrait toutefois poser des problèmes de convergence à cause des nombreuses cellules vides ($n_j = 0$) prévues dans le modèle de Bryant *et coll.* Nous analyserons éventuellement ces problèmes de convergence et les propriétés conditionnelles de l'estimateur par quotient (4.4) dans un document ultérieur.

Si les moyennes de population X_{ij} d'une variable concomitante x sont connues pour toutes les strates, il est alors possible, comme à la section 3.1, d'ajuster un modèle en fonction des moyennes de strates observées y_{ij} . Par exemple, le modèle $y_{ij} = \beta_j x_{ij} + b_j + t_i + \epsilon_{ij}$ où b_j et t_i sont les effets aléatoires et ϵ_{ij} est la moyenne de l'échantillon des erreurs ϵ_{ijk} dans la (i, j) -ième cellule, pourrait être un modèle acceptable. À cela pourrait s'ajouter un élément prédictif de X_{ij} pour les strates non échantillonnées $\beta_j x_{ij} + b_j + t_i$, grâce auquel on pourrait déterminer un estimateur de X . Cette méthode s'apparente aux méthodes d'estimation pour les petites régions, sauf qu'ici le paramètre d'intérêt est la moyenne globale \bar{X} plutôt que les moyennes des cellules X_{ij} . Nous prévoyons analyser les propriétés conditionnelles d'autres estimateurs de \bar{X} dans un document ultérieur.

5. NON-RÉPONSE

5.1 Modèle simplifié

Supposons qu'on obtient m réponses dans un échantillon aléatoire simple de taille n . Définissons W_1 comme la proportion correspondant à la strate de réponse et Y_1 comme la moyenne de la population, où Y_1 et Y_2 sont les moyennes se rapportant respectivement à la strate de réponse et à la strate de non-réponse; de plus, $W_2 = 1 - W_1$. Dans le présent modèle, il n'est pas dit que l'inférence doit reposer sur la valeur observée de m puisque la distribution de m dépend de la valeur inconnue W_1 , laquelle entre dans le calcul du paramètre d'intérêt. En outre, la moyenne (y^m) de l'échantillon des

stratifié. L'estimateur y_U est conditionnellement biaisé:

$$E(y_U) = \sum \sum \left(\frac{n}{n_U G_U} \right) Y_U \neq \sum \sum W_U Y_U = Y;$$

souignons que $E(y_U) = Y_U$ si $n_U > 0$. Ce dernier estimateur présente les mêmes faiblesses que l'estimateur y_D défini dans la section précédente; la difficulté peut être contournée par l'emploi de l'estimateur par quotient

$$y_r = \frac{y_U}{Y_U} = \frac{\sum \sum n_U G_U}{\sum \sum n_U G_U^2} \tag{4.2}$$

où $a_U = \sum \sum n_U G_U / n$, y_r L'estimateur (4.2) est aussi entaché d'un biais conditionnel mais celui-ci est approximativement égal à zéro si $Y_U = Y$ pour tous les (i, j) . Cette dernière condition peut toutefois être invraisemblable dans le cas qui nous occupe puisque les strates sont définies différemment dans le plan de sondage.

Comme à la section 3.1, il semble nécessaire d'utiliser un modèle qui met en relation les strates échantillonnées et les strates non échantillonnées. Vu l'absence de renseignements concomitants, nous pouvons raisonnablement supposer le modèle suivant:

$$y_{ijk} = \mu + \beta_j + \tau_i + \epsilon_{ijk} \tag{4.3}$$

où y_{ijk} est la k -ième observation de la (i, j) -ième cellule, β_j et τ_i sont les effets permanents et ϵ_{ijk} sont les erreurs indépendantes ayant zéro pour moyenne et σ^2 pour variance commune. Malheureusement, les données de l'échantillon ne permettent pas d'estimer la combinaison linéaire $\mu + \beta_j + \tau_i$ pour les strates non échantillonnées et il est, par conséquent, impossible d'estimer les valeurs Y_U correspondantes. On peut remédier à cette situation en supposant que β_j et τ_i sont des variables aléatoires, ce qui nous permet d'obtenir un élément prédictif $\hat{\mu} + \hat{\beta}_j + \hat{\tau}_i$. Dans le cas qui nous occupe, toutefois, le modèle des effets aléatoires peut être moins réaliste que celui qui est défini par l'équation (4.3).

Le défi que ces problèmes posaient a poussé Bankier (1985) à élaborer une méthode itérative appliquée à des échantillons stratifiés indépendants suivant deux critères de stratification distincts. Son estimateur n'est à peu près pas biaisé en fonction du modèle des effets permanents (4.3) tandis que l'estimateur habituel d'Horvitz-Thompson et son équivalent sous forme de quotient le sont.

La méthode de Bankier peut être appliquée au problème de la stratification double. D'après cette méthode, l'estimateur par quotient de Y est défini:

$$y(p) = \sum \sum \frac{n}{G_U(p)} y_U \tag{4.4}$$

où y_U est le total de l'échantillon dans la (i, j) -ième cellule ($y_U = 0$ si $n_U = 0$) et $G_U(p)$ représente les valeurs obtenues au cours de la p -ième itération, de telle sorte que:

$$\sum \frac{n}{G_U(p)} n_U = W_U = \sum W_U \tag{4.5}$$

de y^p , en haussant y_i d'une valeur c suffisamment élevée. Par ailleurs, l'estimateur par quotient

$$y^{pd} = \frac{\sum \frac{E(a_i)}{a_i} W_{ji}}{\sum \frac{E(a_j)}{a_j} W_i}, \tag{3.18}$$

le biais est approximativement égal à zéro si $Y_i = Y$ pour tous les i . L'équation suivante définit un autre estimateur par quotient qui est conditionnellement semblable à y^{pd} :

$$y^{rpsd} = \frac{\sum' W_i}{\sum' W_{ji}}, \tag{3.19}$$

Contrairement à y^{pd} , celui-ci est inconditionnellement non convergent. Ainsi, on peut préférer y^{pd} à y^{rpsd} ou à y^p .

Si l'on disposait simultanément de renseignements sur toutes les strates, il serait possible d'ajuster un modèle en fonction des moyennes de strates observées y_i et d'estimer la moyenne de la population des strates ayant $n_i = 0$. Par exemple, s'il existe une relation linéaire entre les moyennes de la population X_i d'une variable concomitante et les valeurs X_i correspondantes, on peut calculer la valeur prévue de X_i à l'aide de la formule $\hat{\alpha} + \beta X_i = y_i^*$, où $\hat{\alpha}$ et $\hat{\beta}$ sont les estimateurs par la méthode des moindres carrés obtenus en minimisant $\sum' (y_i - \alpha - \beta X_i)^2$. L'estimateur de Y est donc défini par

$$y^{ps*} = \sum' W_{ji} + \sum'' W_{ji}^*, \tag{3.20}$$

où \sum'' désigne la sommation sur l'ensemble des strates où $n_i = 0$. Si le modèle ajusté s'avère satisfaisant, cet estimateur devrait présenter de bonnes propriétés conditionnelles. Nous pouvons donc conclure à partir de cette analyse qu'il n'y a pas de solution facile si $n_i = 0$ pour quelques-unes des strates.

4. STRATIFICATION DOUBLE

Les ouvrages de statistique renferment des solutions ingénieuses pour accroître l'efficacité des estimateurs. Bryant *et coll.* (1960) ont proposé un plan de sondage à stratification double, suivant lequel la taille n_j de l'échantillon de certaines strates (cellules) est nulle. Par cette méthode, nous sommes censés pouvoir estimer la moyenne de la population même si la taille globale n de l'échantillon est inférieure au nombre total de strates. A l'aide d'une méthode de répartition proportionnelle des tailles marginales (n_i, n_j) , Bryant *et coll.* ont obtenu une répartition aléatoire n_{ij} telle que $E(n_{ij}) = (n_i n_j)/n = n W_i W_j$, où W_i et W_j représentent respectivement les totaux marginaux des lignes et des colonnes pour les poids des cellules W_{ij} . Ces mêmes auteurs ont proposé l'estimateur

$$y^v = \frac{1}{n} \sum n_{ij} G_{ij} y_{ij}, \tag{4.1}$$

où $G_{ij} = n^2 W_{ij}/(n_i n_j)$ et y_{ij} peut être assimilée à X_{ij} si $n_{ij} = 0$. L'estimateur y^v est inconditionnellement non biaisé. Toutefois, la distribution de n_{ij} est entièrement connue (puisque tous les W_{ij} sont connus) et, par conséquent, l'ensemble de référence approprié est l'ensemble des échantillons à configuration $\{n_{ij}\}$; en d'autres termes, le plan de sondage proposé devrait être considéré, aux fins d'inférence, comme un échantillonnage aléatoire simple

Hidiroglou et Srinath (1981) ont calculé le biais conditionnel de même que les EQM conditionnelle et inconditionnelle de y , y^* et d'autres variantes de y , mais ils n'ont pas comparé les biais conditionnels de y et de y^* comme ci-dessus.

3.2 $n_i = 0$ pour certaines strates i

Si la taille n de l'échantillon est petite ou qu'il y a un trop grand nombre de strates formées a posteriori, n_i pourrait être égal à zéro pour certaines strates i . L'estimateur de la moyenne de l'échantillon stratifié a posteriori (équation 3.2) devient, dans ce cas, :

(3.14)
$$y^{pst} = \sum' W_i y_i,$$

où \sum' représente la sommation sur l'ensemble des strates ayant une valeur n_i non nulle. L'estimateur (3.14) est conditionnellement biaisé :

(3.15)
$$E_z(y^{pst}) = \sum' W_i X_i \neq \sum W_i X_i.$$

Il demeure conditionnellement biaisé même suivant l'hypothèse idéale que $X_i = Y$ pour toutes les i , qui montre, par ailleurs, que y^{pst} pourrait produire une sous-estimation importante. Cet estimateur est aussi biaisé inconditionnellement. Pour surmonter ces difficultés, il existe une méthode courante qui consiste à combiner des strates semblables pour faire en sorte que $n_i > 0$ pour toutes les i dans l'ensemble réduit de strates. Fuller (1966) propose une solution plus efficace pour le cas particulier où $k = 2$ strates formées a posteriori mais où son cadre d'analyse est inconditionnel en ce sens qu'il tient compte de la probabilité P_1^* , que n_1 soit égal à 0 étant donné que $n_1 = 0$ ou $n_2 = 0$. L'estimateur proposé par Fuller prend la forme suivante :

(3.16)
$$y^F = \frac{W_1 P_1^*}{W_1} y_1 \text{ si } n_2 = 0$$
$$= \frac{W_2 P_2^*}{W_2} y_2 \text{ si } n_1 = 0,$$

où $P_2^* = 1 - P_1^*$. L'estimateur y^F est conditionnellement non biaisé si $n_1 = 0$ ou $n_2 = 0$, mais il est conditionnellement biaisé si nous avons (n_1, n_2) , même lorsque $X_1 = Y_2 = Y$. L'équation suivante définit un estimateur inconditionnellement non biaisé :

(3.17)
$$y_D = \sum \frac{a_i}{W_i} W_i y_i$$

(Doss et coll. 1979), où $a_i = 1$ si au moins une unité de la strate i est incluse dans l'échantillon, $= 0$ dans le cas contraire, et y_i est défini comme X_i si $n_i = 0$ (à noter que $a y_i = 0$ si $n_i = 0$ même si l'on ne connaît pas X_i). Toutefois, l'estimateur y_D , est conditionnellement biaisé puisque

$$E_z(y_D) = \sum' \frac{W_i X_i}{W_i} \neq \sum W_i X_i = Y.$$

Il demeure conditionnellement biaisé même si $X_i = Y$ pour tous les i . Doss et coll. ont critiqué l'efficacité de y_D parce qu'il n'est pas invariant par translation (c'est-à-dire que y_D n'augmente pas d'une valeur c lorsque chaque y_i devient $y_i + c$, c étant une constante arbitraire), et que, par conséquent, on peut accroître indûment la variance

où $W_1^1 = w_1$ si $W_1^* \leq w_1 \leq W_1^{**}$, $= W_1^*$ si $w_1 < W_1^*$, $= W_1^{**}$ si $w_1 > W_1^{**}$ et $W_2^1 = 1 - W_1^1$. Même biaisés, l'estimateur y^{pst} et son équivalent sous forme de quotient devraient être plus efficaces conditionnellement, que y et y_1 si (n_1, n_2) sont donnés. En l'absence de toute condition, l'EQM de y^{pst} devrait être moindre que l'EQM de y , pourvu que $W_1^* \leq W_1 \leq W_1^{**}$. On pourrait aussi recourir à une méthode bayésienne pour estimer W_1 en définissant une distribution préalable pour W_1 .

Exemple 2 (observations aberrantes). L'estimation d'une moyenne de population Y en présence d'observations aberrantes pose un problème comparable à celui de l'exemple précédent concernant les hôpitaux. Supposons qu'il est démontré que la population renferme une faible proportion W_2 d'observations aberrantes (valeurs excessives) mais que W_2 est inconnu, c'est-à-dire $W_1 \gg W_2$ et $Y_2 \gg Y_1$. Alors, si l'échantillon observé ne contient aucune observation aberrante (c'est-à-dire, $w_2 = 0$), nous dirons que y ne reflète aucune-ment la vraie valeur de Y (Chinnappa, 1976) malgré que y soit (inconditionnellement) non biaisé. Le sens de cet énoncé découle du fait que $E_2(y) = Y_1 \ll Y$, où E_2 désigne, comme précédemment, l'espérance conditionnelle.

Par ailleurs, nous considérons y comme une surestimation importante si l'échantillon contenait des observations aberrantes. L'équation (3.8) nous permet d'avancer cela en soulignant que $w_2 \gg W_2$ (puisque W_2 est très petit). Par exemple, si $N_2 = 1$, nous avons $w_2 = 1/n \gg W_2 = 1/N$. Dans ce cas, il nous faut modifier l'estimation y en réduisant le poids associé aux observations aberrantes comprises dans l'échantillon. Nous pourrions, par exemple, faire passer ce poids y de $1/n$ à $1/N$ et rajuster le poids associé aux autres observations de façon que la somme des n poids soit égale à 1 :

$$y^* = \frac{N}{N - n_2} y_1 + \frac{N}{n_2} y_2. \quad (3.12)$$

Le biais relatif conditionnel de y^* est calculé à l'aide de l'équation suivante:

$$\frac{B_2(y^*)}{Y_2} = \left(w_2 \frac{N}{n} - W_2 \right) \delta, \quad (3.13)$$

alors que $B_2(y)/Y_2 = (w_2 - W_2)\delta$. Si $w_2 \frac{N}{n} - W_2 < 0$, alors

$$\left| w_2 \frac{N}{n} - W_2 \right| = W_2 \frac{N}{n} - w_2 \text{ si } 2W_2 < w_2 < W_2 \left(1 + \frac{N}{n} \right).$$

L'inéquation $2W_2 < w_2 (1 + n/N)$ devrait être satisfaite puisque $w_2 \gg W_2$. Si $w_2 n/N - W_2 > 0$, alors

$$\left| w_2 \frac{N}{n} - W_2 \right| = w_2 \frac{N}{n} - W_2 < w_2 - W_2.$$

Par conséquent, le biais conditionnel de l'estimateur y^* devrait être moindre, en valeur absolue, que celui de l'estimateur y .

L'estimateur y^* découle essentiellement de l'estimateur de la moyenne de l'échantillon stratifié a posteriori y^{pst} lorsqu'on pose $N_2 = n_2$. On peut obtenir une solution plus satisfaisante en recueillant, au préalable, des renseignements convenables sur $W_1 (= 1 - W_2)$ à partir des données du recensement par exemple, et en utilisant par la suite l'estimateur y^{pst} ou l'estimateur fondé sur un estimateur de Bayes de W_1 .

supposer l'existence d'un modèle. Il est possible d'exprimer le rapport du biais conditionnel de y à la moyenne de la population des grands hôpitaux, \bar{Y}_2 , comme suit:

$$(3.8) \quad \frac{B_2(y)}{\bar{Y}_2} = (w_1 - w_1\delta) = (w_2 - w_2\delta),$$

où $B_2(y) = E_2(y) - \bar{Y}$ désigne le biais conditionnel de y , $\delta = (\bar{Y}_2 - \bar{Y}_1)/\bar{Y}_2$ et $0 < \delta < 1$ puisque la moyenne de la population \bar{Y}_1 , des petits hôpitaux est moindre que \bar{Y}_2 . Si $w_1 = 1$ (c'est-à-dire que tous les petits hôpitaux sont inclus dans l'échantillon), alors $E_2(y) = \bar{Y}_1 < \bar{Y}$ et y constitue donc une sérieuse sous-estimation. De même, si $w_1 \gg w_2$ (c'est-à-dire que la plupart des petits hôpitaux sont inclus dans l'échantillon), l'équation (3.8) nous amène à déduire que y donnerait lieu à une sérieuse sous-estimation.

Dans l'exemple ci-dessus, il faudrait utiliser l'estimateur de la moyenne de l'échantillon stratifié a posteriori $y_{pst} = w_1\bar{Y}_1 + w_2\bar{Y}_2$, qui est conditionnellement non biaisé sauf si $n_1 = 0$ ou $n_2 = 0$. En fait, il serait préférable d'utiliser un estimateur par quotient

$$(3.9) \quad y_{pst,r} = \frac{y_{pst}}{\bar{X}},$$

où $x_{pst} = w_1x_1 + w_2x_2$ et x_i est la moyenne de l'échantillon de x dans la i ème strate. L'estimateur (3.9) est, à peu de choses près, conditionnellement non biaisé et plus efficace que y_{pst} si n est grand.

Remarque 1. Dans son exemple, Royall devrait en fait utiliser un plan de sondage plus efficace que l'échantillonnage aléatoire simple puisque toutes les valeurs x de la population sont connues; ce pourrait être, par exemple, un échantillonnage aléatoire stratifié suivant x ou, peut-être, une répartition optimale fondée sur les valeurs x .

Remarque 2. Royall justifie l'utilisation de l'estimateur par quotient habituel $y_r = (y/x)\bar{X}$ dans son modèle défini par l'équation (3.6); l'utilisation de cet estimateur ne peut toutefois être justifiée dans le cadre d'analyse conditionnel (échantillonnage répété) puisque y_r est conditionnellement biaisé:

$$(3.10) \quad B_2(y_r) = \bar{X} \left[\frac{w_2\bar{X}_1 + w_2\bar{X}_2}{w_1\bar{X}_1 + w_2\bar{X}_2} - R \right], \quad R = \frac{\bar{X}}{\bar{Y}} \neq 0$$

sauf si $y_1/x_1 = y_2/x_2 = R$. Dans le cas extrême où $w_1 = 1$, $B_2(y_r) = \bar{X}(R_1 - R)$ où $R_1 = \bar{Y}_1/\bar{X}_1$. Ainsi, $B_2(y_r) \leq 0$ dans la mesure où $R_1 \leq R$.

Remarque 3. Si le poids w_1 n'est pas connu mais que \bar{X} l'est, nous ne pouvons utiliser y_{pst} ni $y_{pst,r}$. Royall propose d'utiliser y_r en faisant reposer l'inférence sur la moyenne observée \bar{x} . Toutefois, le choix de \bar{x} est quelque peu arbitraire, et il se pourrait que le biais conditionnel de y_r soit assez important à moins que le modèle décrit par l'équation (3.6) ne soit vrai, du moins approximativement.

Si nous disposons a priori de renseignements valables sur w_1 , par exemple $w_1^* \leq w_1 \leq w_1^{**}$ où w_1^* et w_1^{**} sont connus, on pourrait alors utiliser le pseudo estimateur de la moyenne de l'échantillon stratifié a posteriori: \bar{Y} :

$$(3.11) \quad y_{pst}^* = w_1^*y_1 + w_2^*y_2,$$

pourvu que *toute les valeurs* $n_i \geq 2$, où $(n_i - 1)s_{iv}^2 = \sum_{j \in S_i} (y_{ij} - y_i)^2$ (Holt et Smith 1979). L'intervalle de confiance correspondant, $I_{pst} : y_{pst} \pm z_{\alpha/2} \sqrt{V(y_{pst})}$, est conditionnellement juste. Un autre estimateur de la variance

$$v^*(y_{pst}) = \sum W_i^2 \left[E \left(\frac{1}{n_i} \right) - \frac{1}{N} \right] s_{iv}^2 \quad (3.4)$$

$$= \left(\frac{1}{N} - \frac{1}{N} \right) \sum W_i^2 s_{iv}^2$$

est conditionnellement biaisé sauf lorsqu'on prend la moyenne sur l'ensemble de l'espace échantillon S (en supposant que $P(n_i \leq 1)$ soit négligeable). L'efficacité conditionnelle de l'intervalle de confiance fondé sur (3.4) dépend évidemment de la mesure où les valeurs $1/n_i$ observées divergent de leur espérance respective $E(1/n_i)$. Il convient de souligner que l'intervalle I_{pst} est également juste en l'absence de toute condition, pourvu que $P(n_i \leq 1)$ soit négligeable pour tous les i .

Si $n_i = 1$ pour certaines valeurs de i , on ne peut obtenir d'estimateur de variance conditionnellement non biaisé. Dans ce cas, toutefois, il suffirait d'utiliser la méthode des strates combinées ou la solution modèle proposée à l'origine par Hartley et coll. (1969) pour estimer la variance dans un échantillonnage aléatoire stratifié donnant une unité par strate. Des études empiriques pourraient probablement éclaircir la question de l'applicabilité de ces méthodes. L'argument invoqué habituellement en faveur de y_{pst} de préférence à y est que la variance inconditionnelle de y_{pst} est approximativement égale à la variance obtenue dans la répartition proportionnelle et, de ce fait, est moindre que la variance inconditionnelle de y . Il faut également se rappeler que la répartition proportionnelle est susceptible de produire des gains d'efficacité plutôt modestes. Il est toutefois plus important de noter que la moyenne de l'échantillon (y) est conditionnellement biaisée:

$$E_2(y) = \sum w_i y_i \neq \sum W_i Y_i = Y, \quad w_i = \frac{n_i}{n} \quad (3.5)$$

et que, par conséquent, les inférences correspondantes pourraient être conditionnellement inexactes.

Exemple 1. Supposons $k = 2$ (par exemple, des strates d'hommes et de femmes avec des poids projetés du recensement, W_1 et $W_2 = 1 - W_1$, ou des petits et des grands hôpitaux (Royall 1970)). Royall a utilisé un modèle de superpopulation

$$E_m(y_i) = \beta x_i, \quad i = 1, \dots, N, \quad \beta > 0, \quad x_i > 0 \quad (3.6)$$

pour démontrer que y est conditionnellement biaisé en fonction du modèle. Dans cette équation, E_m désigne l'espérance du modèle, c'est-à-dire,

$$E_m(y) = \beta \bar{x} \neq E_m(Y) = \beta \bar{X} \quad (3.7)$$

à moins que la moyenne de l'échantillon \bar{x} ne corresponde à la moyenne de la population \bar{X} . Dans le modèle de Royall x_i = représente le nombre de lits dans le i ème hôpital et y_i = le nombre de lits occupés dans le i ème hôpital; de plus, x_1, \dots, x_N sont connus. Royall soutient que y produit une sérieuse sous-estimation si l'échantillon observé contient tous (ou presque tous) les petits hôpitaux puisque $B_m(y) = E_m(y) - E_m(Y) = \beta(\bar{x} - \bar{X})$ et $\bar{x} \ll \bar{X}$. Nous pouvons démontrer la même chose dans notre cadre d'analyse conditionnel sans

Conditionnellement, le seuil de confiance de I_p est environ $1 - \alpha$ si la valeur de v est élevée. Un autre estimateur de la variance, soit

$$(2.6) \quad v^*(v_p) = \left[E\left(\frac{1}{v}\right) - \frac{1}{N} \right] s_{vy}^2$$

est conditionnellement biaisé sauf lorsqu'il s'étend à l'ensemble de l'espace échantillon S . Les équations (2.4) et (2.6) permettent de déduire que $v(v_p) < v^*(v_p)$ si $1/v < E(1/v)$, et l'inverse si $1/v > E(1/v)$. Par conséquent, l'intervalle de confiance fondé sur (2.6) serait trop étroit si $E(1/v) < 1/v$ et correspondrait alors à un seuil de confiance inférieur à $1 - \alpha$, et serait trop étendu si $E(1/v) > 1/v$, ce qui donnerait un seuil de confiance supérieur à $1 - \alpha$. Il convient de souligner qu'un intervalle conditionnellement juste est également juste en l'absence de toute condition.

3. ÉCHANTILLONNAGE ALÉATOIRE SIMPLE SANS REMISE

Supposons qu'un échantillon aléatoire simple de taille déterminée n est prélevé sans remise. En l'absence de sous-ensembles identifiabiles, l'ensemble de référence approprié est l'ensemble S constitué de $\binom{N}{n}$ échantillons s , de taille n , la moyenne de l'échantillon y_n est non biaisée et sa variance est estimée sans distorsion par l'équation

$$(3.1) \quad v(y_n) = \left(\frac{1}{n} - \frac{1}{N} \right) s_{ny}^2$$

où $(n - 1)s_{ny}^2 = \sum_{i \in s} (y_i - y_n)^2$. L'intervalle de confiance correspondant est donné par I_p : $y_n \pm z_{\alpha/2} \sqrt{v(y_n)}$ avec un seuil de confiance d'environ $1 - \alpha$ si n est élevée. Supposons maintenant qu'il existe des sous-ensembles identifiabiles en ce sens que nous observons un échantillon de configuration $\tilde{n} = (n_1, \dots, n_k)$ réparti dans k strates formées à posteriori et ayant des poids donnés $W_i = N_i/N$. Idéalement, on aurait dû procéder par échantillonnage stratifié, mais il n'existait pas de base de stratification. L'ensemble de référence approprié est désormais l'ensemble $S_{\tilde{n}}$, composé de $\prod \binom{N_i}{n_i}$ échantillons de configuration réalisée \tilde{n} , la distribution de celle-ci étant entièrement connue.

3.1 Tous les $n_i \geq 1$

Si toutes les valeurs de $n_i \geq 1$ observées sont supérieures ou égales à 1, l'estimateur habituel de la moyenne de l'échantillon stratifié à posteriori

$$(3.2) \quad y^{pst} = \sum W_i y_i$$

est conditionnellement non biaisé étant donné \tilde{n} puisque $P(s|\tilde{n}) = \prod \binom{N_i}{n_i}^{-1}$, en d'autres termes, l'échantillon observé s est, conditionnellement, un échantillon aléatoire stratififié (s_1, \dots, s_k) avec des strates de taille n_i . Dans l'équation ci-dessus, y_i désigne la moyenne de l'échantillon dans la i ème strate. La variance conditionnelle $V_2(y^{pst})$ offre une mesure d'incertitude convenable; elle est estimée sans biais par l'équation

$$(3.3) \quad v(y^{pst}) = \sum W_i^2 \left(\frac{1}{n_i} - \frac{1}{N} \right) s_{iy}^2$$

les valeurs 1, ..., n. Définissons t_i comme le nombre de fois que la ième unité de la population Y sont tion est incluse dans s . Alors, deux estimateurs courants de la moyenne de population \bar{Y} sont donnés par:

(2.1)
$$\bar{y}_n = \frac{1}{n} \sum_{i \in s} t_i y_i$$

la moyenne de l'échantillon fondée sur les n prélèvements, et

(2.2)
$$\bar{y}_v = \frac{1}{v} \sum_{i \in s} y_i$$

la moyenne fondée sur les unités distinctes de s . Les deux moyennes \bar{y}_n et \bar{y}_v sont inconditionnellement non biaisées selon l'ensemble de référence S , et la variance inconditionnelle de \bar{y}_v est toujours moindre que celle de \bar{y}_n . Ainsi, pour des raisons d'efficacité, on devrait préférer \bar{y}_v à \bar{y}_n . L'estimateur d'Horvitz-Thompson

(2.3)
$$y_{HT} = \frac{1}{N} \sum_{i \in s} \pi_i = \frac{1}{v} \frac{E(v)}{v}$$

est aussi inconditionnellement non biaisé; dans l'équation ci-dessus, π_i désigne la probabilité que l'unité i soit incluse au moins une fois dans l'échantillon:

$$\pi_i = \frac{E(v)}{N} = 1 - \left(1 - \frac{1}{N}\right)^n$$

La comparaison des variances de \bar{y}_v et de y_{HT} indique que \bar{y}_v n'est pas toujours supérieur à y_{HT} .

À la lumière des affirmations de Durbin (1969), il est clair que l'on devrait faire reposer l'inférence sur la valeur observée de v , en d'autres mots, l'ensemble de référence approprié est l'ensemble S_v constitué de $\binom{N}{v}$ échantillons de taille réelle v , et non l'ensemble S . Heureusement, l'ensemble se prête facilement à l'inférence conditionnelle puisque $P(s_v | v) = \binom{N}{v}^{-1}$; en d'autres termes, l'échantillon d'unités distinctes observé s_v , est, conditionnellement, un échantillon aléatoire simple de taille v prélevé sans remise. Il s'ensuit que \bar{y}_v est conditionnellement non biaisé, c'est-à-dire que $E_2(\bar{y}_v) = \bar{Y}$ où E_2 représente l'espérance conditionnelle, tandis que $E_2(y_{HT}) = [v/E(v)] \bar{Y} \neq \bar{Y}$ de sorte que y_{HT} est conditionnellement biaisé. Il faudrait donc préférer \bar{y}_v à y_{HT} , en dépit de la comparaison peu concluante de leurs variances inconditionnelles. Soulignons que y_{HT} constituerait une sous-estimation marquée si la valeur v observée était de beaucoup inférieure à $E(v)$. La variance conditionnelle $V_2(\bar{y}_v)$, est une mesure d'incertitude convenable; elle est estimée sans distorsion par la formule suivante:

(2.4)
$$v(y_v) = \left(\frac{1}{v} - \frac{1}{N}\right) s_{y_v}^2$$

où $(v - 1)s_{y_v}^2 = \sum_{i \in s} (v_i - \bar{y}_v)^2$ et V_2 désigne la variance conditionnelle. L'intervalle de confiance approprié pour \bar{Y} est obtenu par la formule suivante:

(2.5)
$$I_v = \bar{y}_v \pm z_{\alpha/2} \sqrt{v(y_v)}.$$

celui-ci contient des sous-ensembles identifiables. Nous illustrerons plus loin la notion de sous-ensemble identifiable par des exemples d'échantillons aléatoires de tailles diverses. Le choix de l'ensemble de référence approprié n'est toutefois pas unique. En fait, on pourrait concrètement considérer l'échantillon prélevé s comme unique, mais alors aucune inférence ne serait possible dans le cadre d'un échantillonnage répété puisque l'ensemble de référence approprié serait un singleton (Holt et Smith 1979).

L'inférence conditionnelle est un sujet qui a été très discuté en statistique classique depuis Fisher (1925). En testant, par exemple, la notion d'indépendance dans un tableau de contingence 2×2 , Fisher prétendait que l'inférence devait dépendre des totaux marginaux observés des lignes et les colonnes même si les marges n'étaient pas déterminées par le plan de sondage. Yates (1984) a réexaminé ce problème. Le choix d'un ensemble de référence approprié ne s'impose pas toujours à l'évidence, mais il paraît raisonnable d'appliquer les règles suivantes : 1) Il faut choisir une méthode conditionnelle *avant* l'observation des données, surtout dans le domaine public. 2) Il faut répartir conditionnellement l'ensemble S de telle manière que les sous-ensembles obtenus ne renferment pas ou à peu près pas de renseignements sur les paramètres d'intérêt; en d'autres mots, les statistiques servant d'indices aux sous-ensembles devraient être des fonctions auxiliaires des observations (Cox et Hinkley 1974, p. 38). 3) Si la taille des échantillons est aléatoire (par exemple, les tailles d'échantillons de domaines) et que la distribution de la population est entièrement (ou du moins partiellement) connue, les inférences dépendront des tailles d'échantillons observées. Dans ce contexte, Durbin (1969, p. 643) affirme que si la taille de l'échantillon est déterminée de manière aléatoire et qu'il se trouve qu'on obtienne un grand échantillon, on sait parfaitement bien que les mesures recherchées sont plus précises que s'il s'agissait d'un échantillon de petite taille. Il précise que, de toute évidence, on devrait se servir des renseignements connus sur la taille de l'échantillon pour interpréter les résultats. Selon Durbin, du point de vue de l'analyse des données effectivement observées, il semble tout à fait incorrect de faire une moyenne des variations hypothétiques de la taille de l'échantillon lorsqu'en fait, on connaît exactement la taille de l'échantillon.

La présente analyse se limitera à l'inférence conditionnelle en présence de tailles d'échantillon aléatoires, selon la règle 3 énoncée ci-dessus. Malgré cette restriction, nous montrerons qu'il n'est pas toujours facile, en pratique, de faire des inférences conditionnelles. Nous débutons notre analyse avec des exemples simples, puis nous poursuivons avec des problèmes plus complexes. Pour les enquêtes par sondage, ce sont Holt et Smith (1979) qui avancent les arguments les plus convaincants en faveur de l'inférence conditionnelle, bien que leur analyse soit limitée à la stratification a posteriori d'un échantillon aléatoire simple (EAS) (voir section 3.1). Lahiri (1969) a souligné la difficulté de convaincre les utilisateurs de données statistiques, personnes intelligentes mais peu versées en la matière, de l'importance réelle des estimations provenant des enquêtes par sondage. Il souligne, notamment, l'aberration qui consiste à se servir implicitement de l'erreur type (d'échantillonnage) comme mesure de précision de l'estimation observée (de l'échantillon), en s'inspirant d'un certain nombre d'exemples tirés de la théorie actuelle.

2. ÉCHANTILLONNAGE ALÉATOIRE SIMPLE AVEC REMISE

L'échantillonnage aléatoire simple (EAS) avec remise est rarement utilisé en pratique, mais il offre un moyen facile de s'initier à l'inférence conditionnelle. Supposons qu'un échantillon aléatoire simple s , de taille n est prélevé avec remise dans une population de taille N de telle sorte que l'ensemble S contient N^n échantillons s . Soit ν le nombre d'unités distinctes dans s . Alors, ν est une variable aléatoire pouvant prendre

Inference conditionnelle dans les enquêtes par sondage

J.N.K. RAO¹

RÉSUMÉ

L'auteur jette un regard critique sur les méthodes traditionnelles d'inference dans les enquêtes par sondage. Il souligne la nécessité de faire reposer l'inference sur des sous-ensembles identifiables de la population. Utilisant des échantillons aléatoires de tailles diverses, il apporte un certain nombre d'exemples concrets d'inferences dépendant de la configuration réelle de l'échantillon ainsi que les problèmes qui peuvent y être associés. Parmi les exemples choisis, mentionnons l'estimation d'une moyenne de population suivant un échantillonnage aléatoire simple, l'estimation d'une moyenne de population en présence d'observations aberrantes, l'estimation du total et de la moyenne d'un domaine, l'estimation de la moyenne d'une population dans le contexte d'une stratification double et l'estimation d'une moyenne de population suivant des plans de sondage généraux. Enfin, l'auteur analyse le biais conditionnel et la variance conditionnelle des estimateurs de la moyenne d'une population (de la moyenne ou du total d'un domaine) de même que les intervalles de confiance correspondants.

MOTS CLÉS : Inference conditionnelle, biais conditionnel, variance conditionnelle, moyenne de population, taille d'échantillons aléatoires.

1. INTRODUCTION

D'après la théorie classique de l'inference dans les enquêtes par sondage, le plan de sondage définit l'espace échantillon S (ensemble des échantillons possibles s) et les probabilités de sélection correspondantes $p(s)$. Le choix d'un estimateur repose sur le critère de convergence ou d'absence de biais et sur la comparaison des erreurs quadratiques moyennes (EQM), suivant un échantillonnage répété avec des probabilités $p(s)$, l'espace échantillon S étant utilisé comme ensemble de référence. Ainsi, l'estimateur \hat{Y} d'une moyenne de population Y est sans biais si $E(\hat{Y}) = \sum_{s \in S} p(s) \hat{Y}_s = Y$, où \hat{Y}_s est la valeur de \hat{Y} pour l'échantillon s . L'EQM de l'estimateur \hat{Y} est donnée par $EQM(\hat{Y}) = \sum_{s \in S} p(s) (\hat{Y}_s - Y)^2$, et \hat{Y} est convergent si son EQM tend vers zéro au fur et à mesure que la taille de l'échantillon augmente. Un estimateur convergent ou sans biais de $EQM(\hat{Y})$, désigné par $eqm(\hat{Y})$, renferme une mesure d'incertitude pour la valeur de \hat{Y} . Si \hat{Y} est sans biais ou convergent, on obtient pour les valeurs observées \hat{Y}_s et $eqm(\hat{Y}_s)$ un intervalle de confiance de grand échantillon défini par l'équation suivante (à un seuil de confiance $1 - \alpha$):

$$I_s = \hat{Y}_s \pm z_{\alpha/2} \sqrt{eqm(\hat{Y}_s)}, \quad (1)$$

où $z_{\alpha/2}$ est la borne supérieure $\alpha/2$ -de l'intervalle de confiance pour une variable $N(0, 1)$. L'équation (1) signifie que, dans un échantillonnage répété ayant comme ensemble de référence S , environ 100 $(1 - \alpha)\%$ des intervalles I_s renfermeront la vraie valeur Y . Il est juste de comparer des erreurs quadratiques moyennes inconditionnelles $EQM(\hat{Y})$, au moment de l'élaboration de l'enquête, mais il se peut que l'espace échantillon S ne soit plus l'ensemble de référence qui convienne à l'inference une fois l'échantillon s prélevé, si

¹ J.N.K. Rao, département de mathématiques et de statistiques, Université Carleton, Ottawa (Ontario), Canada K1S 5B6.

ANNEXE

Calcul de la valeur implicite z_i pour l'imputation dans les cas de non-réponse au questionnaire

Pour les cas c) et d) du Tableau 6, l'estimation au niveau de la cellule b est définie:

$$\bar{Y}_b = \bar{F}_{bQV}(wa)_i + (\bar{F}_b - \bar{F}_{bQV}) \tag{A.1}$$

$$= \bar{F}_b + [(wa)_i - 1] \bar{F}_{bQV}$$

Pour ce qui est de c), $(wa)_i - 1 = (n_b - m_b)/m_{bQ}$

$$= \sum t_i(1 - \delta_i)/m_{bQ}$$

$$\text{ou } \bar{Y}_b = \bar{F}_b + \sum t_i \pi_i^{-1} (1 - \delta_i) \bar{F}_{bQV} / \pi_i^{-1} m_{bQ} \tag{A.2}$$

par ailleurs, en mettant (A.2) en équation avec (A.1), compte tenu de la définition de \bar{F}_b donnée en (3.10), on peut voir que la valeur imputée z_i est défini par \bar{Y} , ce qui correspond à l'élément c) du Tableau 6.

De même, lorsque des taux de réponse pondérés sont utilisés, on peut voir que la valeur imputée implicite z_i équivaut à $\bar{F}_{bQV}/\bar{M}_{bQ}$, ce qui correspond à l'élément d) du Tableau 6. On obtient les éléments a) et b) du Tableau 6 en posant $m_{bQ} = m_b$ and $\bar{M}_{bQ} = \bar{M}_b$.

BIBLIOGRAPHIE

HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

LESSLER, J.T. (1979). An expanded survey error model. Dans *Incomplete Data in Sample Surveys*, volume n° 3 - Compte rendu des travaux du symposium (éd. W.G. Madow, I. Olkin, et D.B. Rubin), San Diego: Academic Press, 259-270.

PLATEK, R. (1977). Some factors affecting nonresponse. *Techniques d'enquête*, 3, 191-214.

PLATEK, R. (1980). Causes of incomplete data, adjustments and effects. *Techniques d'enquête*, 6, 93-132.

PLATEK, R., et GRAY, G.B. (1978). Non-response and imputation. *Techniques d'enquête*, 4, 144-177.

PLATEK, R., et GRAY, G.B. (1979). Methodology and application of adjustments for nonresponse. Exposé présenté à la 42^e séance de l'Institut internationale de statistique, Manille, Philippines.

PLATEK, R., et GRAY, G.B. (1983). Part V - Imputation Methodology: Total Survey Error. Dans *Incomplete Data in Sample Surveys*, volume n° 2 - Theory and Bibliographies (éd. W.G. Madow, I. Olkin, et D.B. Rubin), San Diego: Academic Press, 249-333.

POTERBA, J.M., et SUMMERS, L.H. (1984). Response Variation in the CPS: Caveats for the unemployment analyst. *Monthly Labour Review*, mars 1984, résumés de recherche, 37-43.

Le facteur de correction de poids indiqué en c) $(n_b - m_b + m_{b0})/m_{b0} = 1 + (n_b - m_{b0})/m_{b0}$ est \geq à celui indiqué en a) (n_b/m_b) puisque $m_{b0} \leq m_b$ (voir Tableau 6). Ainsi, pour un taux de réponse donné m_b/n_b dans la cellule, la variance d'une estimation obtenue à l'aide du facteur de correction c) risque d'être plus élevée que celle d'une estimation obtenue à l'aide du facteur de correction a). L'augmentation de la variance peut être compensée ou non par une réduction possible du biais dû à l'imputation dans l'erreur quadratique moyenne globale. Le même raisonnement, s'appliquant aux taux de réponse pondérés figurant en d) $(N_b - M_b + M_{b0})/M_{b0}$ par rapport à celui figurant en b) (N_b/M_b) puisque $M_{b0} \leq M_b$. Dans l'échantillonnage avec ppt, l'utilisation de taux de réponse pondérés par rapport à des taux de réponse non pondérés nous amène à une autre conclusion intéressante. Platek et Gray (1983, p. 264-265), on montré que lorsque les probabilités de réponse et de sélection, c'est-à-dire, α_i et π_i sont interdépendantes, les facteurs de correction de poids formés avec des taux de réponse pondérés tendent à être plus élevés que ceux formés avec des taux non pondérés. Ainsi, lorsqu'il y a une corrélation positive entre α_i et π_i , $E(N_b/M_b) > E(n_b/m_b)$ parallèlement, $E[(N_b - M_b + M_{b0})/M_{b0}] > E[(n_b - m_b + m_{b0})/m_{b0}]$, où $E = E_1 E_2$, est l'espérance mathématique pour l'ensemble des échantillons d'unités et des sous-échantillons d'unités répondantes (voir Platek et Gray (1983), p. 251).

Bref, quel que soit le facteur de correction de poids utilisé pour compenser la non-réponse au questionnaire, il n'est pas certain que les valeurs imputées implicites z_i se rapprochent des valeurs réelles X_i ou même des valeurs observées probables y_i . Tout ce que nous pouvons faire relativement à cette question, c'est d'espérer que les cellules de correction créées pour compenser la non-réponse au questionnaire permettront d'uniformiser le plus possible les caractéristiques des répondants et des non-répondants dans les cellules. La formation et la délimitation des cellules de correction sont donc fondamentales pour la compensation, peu importe le genre de facteur de correction de poids qui est utilisé.

7. CONCLUSION

Comme nous l'avons vu dans les sections précédentes, il n'y a pas de solution toute faite à la non-réponse, peu importe le genre de non-réponse qui se produit. Il s'agit tout d'abord de réduire au minimum le taux de non-réponse sans que cela n'engendre des frais trop élevés ou ne supprime l'actualité des données d'enquête. Au départ, on devra savoir prévoir les cas de non-réponse et élaborer des stratégies d'imputation. Si la non-réponse se manifeste à peu près de la manière prévue, le traitement des données se déroulera conformément au programme établi, notamment au moyen de substitutions et de corrections de poids appropriées. Il est clair que les enquêtes permanentes ou répétées permettent un déroulement plus ordonné des diverses opérations (collecte des données, publication, etc.) que les enquêtes spéciales ou ponctuelles, où il est difficile pour le concepteur de prévoir tous les obstacles qui peuvent surgir en cours d'enquête, comme les refus inattendus ou le manque d'intérêt aussi bien de la part des intervieweurs que des répondants.

Pour surmonter les problèmes de non-réponse, il est essentiel de faire une étude continue des taux de non-réponse par caractéristique d'enquête (dans le cas de la non-réponse à une question) et des causes de la non-réponse et, si possible, d'analyser, par la même occasion, les probabilités de réponse à une question et au questionnaire afin de pouvoir estimer les biais dus à l'imputation à partir de l'enquête proprement dite. Par ailleurs, il est souhaitable d'analyser le biais dû à l'imputation en approfondissant la recherche sur les estimations fondées sur les modèles et d'améliorer les estimations à l'aide de renseignements additionnels.

pour l'unité non répondante i est indiquée au Tableau 6 pour chacun des 4 facteurs de correction de poids définis ci-dessus. Pour bien comprendre les expressions figurant au Tableau 6, il est nécessaire de définir les termes suivants:

(3.10)
$$T_b = \sum_{i=1}^{N_b} t_i \pi_i^{-1} [\delta_i \delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}] = \text{total pondéré de l'échantillon d'unités ayant répondu entièrement au questionnaire, } y \text{ compris les imputations pour non-réponse à une question mais sans les corrections de poids effectuées au moyen de l'inverse du taux de réponse au questionnaire,}$$

(3.11)
$$T_{by} = \sum_{i=1}^{N_b} t_i \pi_i^{-1} \delta_i \delta_{iy} y_i = \text{total pondéré de l'échantillon d'unités ayant répondu du partiellement ou entièrement au questionnaire en fonction de la caractéristique } y,$$

(3.12)
$$T_{bQy} = \sum_{i=1}^{N_b} t_i \pi_i^{-1} \delta_i \prod_{Q=1}^b \delta_{iQ} y_i = \text{total pondéré de l'échantillon d'unités ayant répondu du partiellement ou entièrement au questionnaire en fonction de la caractéristique } y, \text{ sans les unités de la cellule qui ont fait l'objet d'une imputation pour non-réponse à une question.}$$

Alors, $T_{bQy} \leq T_{by} \leq T_b$.

Tableau 6
Valeur imputée implicite pour les non-répondants au questionnaire, selon le facteur de correction de poids (au niveau de la cellule)

	Facteur de correction de poids	Equation correspondante	Valeur imputée implicite lorsque $i = 0$	Description
a)	n_b/m_b	(3.6) $T_b/(\pi_i^{-1} m_b)$	Taux de réponse au questionnaire non pondéré	Taux de réponse au questionnaire non pondéré
b)	N_b/M_b	(3.8) T_b/M_b	Taux de réponse au questionnaire pondéré	Taux de réponse au questionnaire pondéré
c)	$n_b - m_b + m_{bQ}$	(3.7a) $T_{bQy}/\pi_i^{-1} m_{bQ}$	Taux de réponse au questionnaire non pondéré parmi les unités qui ont répondu à toutes les questions	Taux de réponse au questionnaire non pondéré parmi les unités qui ont répondu à toutes les questions
d)	$N_b - M_b + M_{bQ}$	(3.9) T_{bQy}/M_{bQ}	Taux de réponse au questionnaire pondéré parmi les unités qui ont répondu à toutes les questions	Taux de réponse au questionnaire pondéré parmi les unités qui ont répondu à toutes les questions

Note: Dans le cas d'un échantillon auto-pondéré (par exemple, EAASSR), la valeur imputée implicite z_i devient la moyenne simple des répondants, pour a) et b), et la moyenne simple des répondants (à l'exclusion de ceux qui n'ont pas répondu à toutes les questions) pour c) et d).

* Voir annexe I pour les calculs.

naires qui ne sont pas entièrement remplis probablement des erreurs de réponse et des erreurs d'imputation tandis que les questionnaires entièrement remplis ne renferment que des erreurs de réponse. De plus, en appliquant un facteur de correction de poids plus grand aux unités qui ont répondu à toutes les questions, on peut obtenir des estimations entachées d'une erreur quadratique moyenne plus faible que si l'on appliquait le même facteur de correction de poids à toutes les unités répondantes de la cellule. À notre connaissance, les facteurs de correction de poids que nous venons de décrire n'ont pas été utilisés jusqu'à maintenant, mais il serait peut-être bon que l'on envisage de le faire s'il est démontré que la réduction du biais compense l'augmentation de variance causée par l'existence de poids différents.

En ce qui concerne les unités ayant des probabilités d'échantillonnage différentes, il existe un facteur de correction de poids fondé sur l'échantillon pondéré et les unités répondantes d'une cellule plutôt que sur les échantillons non pondérés. Dans ce cas,

$$(3.8) \quad (w)_i = N_b / M_b,$$

où $M_b = \sum_{i \in b} \pi_i^{-1} t_i \delta_i$ est le nombre pondéré d'unités répondantes dans la cellule b . Pour ce qui a trait aux facteurs de correction de poids définis en (3.7a), que l'on applique uniquement aux unités qui ont répondu entièrement au questionnaire, nous avons

$$(3.9) \quad (w)_i = [N_b - (M_b - M^{b0})] / M^{b0}$$

où $M^{b0} = \sum_{i \in b} \pi_i^{-1} t_i \delta_i \Pi_{\delta_i=1}^{q=1} \delta_{iq}$, est le nombre pondéré d'unités de la cellule b , qui ont répondu à toutes les questions du questionnaire.

$\delta_{iq} = 1$ ou 0 selon que l'unité i a répondu ou non à la question q du questionnaire d'enquête comportant Q questions; par conséquent, $\Pi_{\delta_i=1}^{q=1} \delta_{iq} = 1$ seulement si l'unité i a répondu à toutes les questions du questionnaire.

On pourrait justifier l'utilisation de (3.9) au lieu de (3.8) par les mêmes raisons qui ont motivé l'utilisation de (3.7a) au lieu de (3.6). Il est nécessaire toutefois d'expliquer pourquoi il est préférable d'utiliser les taux de réponse pondérés au lieu des taux de réponse non pondérés; cette explication est fournie après le Tableau 6.

À partir des équations (3.7a) et (3.9), il serait possible de définir une expression $(w)_i$ distincte pour chaque question q ou chaque caractéristique y définie par une ou plusieurs questions. Malheureusement, des questions ou des caractéristiques différentes supposeraient nécessairement des facteurs de correction de poids différents dans une cellule de correction, ce qui créerait des incohérences entre diverses caractéristiques figurant dans des tableaux publiés. Pour assurer l'uniformité des poids de l'enquête et des facteurs de correction de poids, il faudrait que $(w)_i$ dépende uniquement de l'unité et non de la question ou de la caractéristique. L'imputation pourrait néanmoins être permise pour certaines questions mais non pour d'autres, comme les principales questions incluses dans les facteurs de correction de poids (3.7a) ou (3.9), pour autant qu'il y ait cohérence entre les inclusions et les exclusions à l'intérieur de la cellule de correction. Pour calculer un facteur de correction de poids, on pourrait, par exemple, envisager de faire une imputation par déduction logique plutôt que par la méthode *hot deck*, en se fondant sur un questionnaire entièrement rempli.

Pour chacun des facteurs de correction de poids définis ci-dessus (3.6 à 3.9), il est possible de montrer que (2.2) est une forme particulière de (2.1) lorsque z_i correspond à une moyenne pondérée ou non pondérée des répondants. Ainsi, la valeur imputée implicite z_i

Tableau 5
Nombre de cellules de correction, et moyenne et distribution de fréquence des facteurs de correction de poids selon la région et le genre d'unité. Janvier 1983

Région		Nombre		Moyenne		Moyenne		1-		1.01-		1.02-		1.03-		1.04-		1.05-		1.06-		1.07-		1.08-		1.09-		1.10-		1.10 +	
Genre d'unité		de cellules		(wa) _i																											
Atl.	UNAR	254	1.0250	143	6	22	21	13	13	9	7	8	2	10																	
Atl.	UAR	123	1.0246	58	5	11	15	14	4	3	6	4	1	2																	
Qué.	UNAR	126	1.0550	72	2	8	10	10	6	8	6	0	1	3																	
Qué.	UAR	185	1.0265	106	0	7	8	23	11	4	5	7	3	11																	
Ont.	UNAR	120	1.0333	58	1	10	11	11	8	4	2	2	2	11																	
Ont.	UAR	252	1.0416	116	1	13	24	21	16	9	9	8	10	25																	
Pt.	UNAR	328	1.0348	167	5	17	22	23	24	15	12	10	8	25																	
Pt.	UAR	149	1.0306	40	23	23	20	13	8	7	3	5	4	3																	
C.B.	UNAR	85	1.0468	38	3	7	8	8	2	5	1	1	1	11																	
C.B.	UAR	119	1.0412	46	4	7	15	10	7	7	7	3	3	10																	
Can.	UNAR	913	1.0358	478	17	64	72	65	53	41	28	21	14	60																	
Can.	UAR	828	1.0337	366	33	61	82	81	46	30	30	27	21	51																	
Canada		1 741	1.0348	844	50	125	154	146	99	71	58	48	35	111																	

Si on ne connaît pas les caractéristiques des non-répondants, il n'est pas possible de déterminer avec exactitude la valeur maximale que devrait avoir le facteur de correction de poids si l'on veut éviter un biais excessif et une augmentation de la variance attribuable à une taille réelle d'échantillon moindre. Si on fixe arbitrairement cette valeur maximale à 1.05 pour l'EPA (niveau que les spécialistes d'enquête posent parfois par hypothèse), on observe qu'environ le quart des unités de compensation au Canada (441 sur 1741) avaient un facteur de correction de poids de 1.05 ou plus en janvier 1983. Dans beaucoup d'autres enquêtes comme celles portant sur les revenus et les dépenses, le taux de non-réponse est généralement plus élevé et les facteurs de correction de poids de presque toutes les cellules dépasseraient vraisemblablement le seuil critique si celui-ci était maintenu par hypothèse à 1.05.

Il existe d'autres genres de facteurs de correction de poids dans les cellules. Par exemple, on pourrait exclure de la cellule *b*, telle qu'elle est définie plus haut, les unités qui n'ont pas répondu à toutes les questions du questionnaire. Supposons que la cellule *b* compte m_{b0} unités qui ont répondu à toutes les questions du questionnaire. Pour les $(m_b - m_{b0})$ unités répondantes de la cellule qui n'ont pas répondu à toutes les questions, le facteur de correction de poids $(wa)_i = 1$, et pour les m_{b0} unités qui ont répondu à toutes les questions, le facteur de correction de poids est défini par:

$$(wa)_i = [n_b - (m_b - m_{b0})] / m_{b0}, \text{ which exceeds } n_b / m_b. \quad (3.7a)$$

Dans ce qui suit, nous tentons de démontrer qu'il est préférable de ne pas appliquer de facteur de correction de poids, c'est-à-dire, $(wa)_i = 1$, aux unités de la cellule qui n'ont pas répondu à toutes les questions du questionnaire mais d'appliquer un facteur de correction supérieur à n_b / m_b (3.7a) aux unités qui ont répondu à toutes les questions. Les question-

correction. Par conséquent, si la population d'une cellule est composée de N unités et que cette cellule est représentée par un échantillon de n unités, où :

$$n_b = \sum_{i \in b} t_i \text{ la taille de l'échantillon de la cellule } b, \text{ cette valeur pouvant ou non être une constante selon la définition de la cellule;}$$

$$N_b = \sum_{i \in b} \pi_i t_i, \text{ une estimation de la taille de la population de la cellule } b; N_b \text{ ne devrait normalement pas être connue, sauf s'il s'agit d'un recensement;}$$

$$m_b = \sum_{i \in b} t_i \delta_i = \text{nombre d'unités répondantes dans la cellule } b, \text{ c'est-à-dire la taille de l'échantillon répondant,}$$

alors, $(wa)_i = n_b/m_b$ lorsque i appartient à la cellule de correction b . (3.6)

Avant de définir d'autres facteurs de correction de poids, nous allons analyser plus à fond ce facteur fréquemment utilisé qu'est le taux de réponse inverse non pondéré dans une cellule, décrit en (3.6). Avec $(wa)_i = n_b/m_b$, l'estimation du total définie en (2.2) peut être réécrite comme un cas particulier de (2.1), où z_i est défini comme suit:

$$z_i = \mathcal{F}_b / \pi_i^{-1} m_b, \quad (3.7)$$

où $\mathcal{F}_b = \sum_{i \in b} \pi_i^{-1} t_i \delta_i [\delta_{iy} z_i + (1 - \delta_{iy}) z_{iy}]$, le total pondéré de l'échantillon d'unités répondantes dans la cellule b . Lorsqu'une cellule présente des poids d'échantillons égaux, la valeur imputée z_i se réduit à la valeur moyenne des m_b répondants dans la cellule. Si l'on substitue l'expression (3.7) à z_i dans l'équation (2.1), on peut montrer que l'estimation est similaire à celle obtenue dans l'équation (2.2) si l'on considère que $(wa)_i = (n_b/m_b)$. On peut, par conséquent, considérer l'imputation pour la non-réponse au questionnaire comme une substitution de $z_i = \mathcal{F}_b / (\pi_i^{-1} m_b)$ dans l'équation (2.1) ou comme une correction des poids de l'échantillon dans l'équation (2.2) au moyen du facteur $(wa)_i = n_b/m_b$. En ce qui concerne la correction de poids, on poserait $z_i = 0$ dans l'expression (3.4), ce qui aurait pour effet de maintenir les cinq éléments de \mathcal{F}_b . Par ailleurs, on pourrait utiliser la valeur imputée z_i telle qu'elle est définie dans l'équation (2.1) et on poserait alors $(wa)_i = 1$ dans l'expression (3.5), ce qui rendrait nul le dernier élément de \mathcal{F}_b . Par conséquent, si l'on veut analyser l'effet de la correction de poids $((wa)_i > 1)$, il faut considérer dans une même bloc l'élément négatif (3.4) et l'élément positif (3.5); mais si l'on veut analyser l'effet de la valeur imputée implicite z_i , définie en (3.7), seul l'élément (3.4) peut être pris en considération. Le facteur de correction de poids (n_b/m_b) est utilisé dans l'EPA. Les cellules de correction de l'EPA sont des U.P.E. liées au plan de sondage dans les unités non autoréprésentatives (UNAR) et des strates (sous-unités) d'îlots urbains contigus dans les unités autoréprésentatives (UAR). Le tableau 5 donne, par région et par genre d'unité, le nombre de cellules, la moyenne non pondérée des facteurs de correction de poids et leur distribution de fréquence dans les intervalles 1-1.01, 1.01-1.02, ..., 1.10 et plus pour l'enquête de janvier 1983. Le facteur de correction de poids moyen pour le Canada (1.0348) est inférieur à ce qu'on s'attendrait d'obtenir avec un taux de non-réponse d'environ 5%. Cette moyenne apparentement faible s'explique par le fait que certains non-répondants, qui ont toutefois répondu au questionnaire du mois précédent, sont considérés comme des répondants aux fins du calcul du taux de réponse inverse. Cette mesure vise de 20 à 30% des non-répondants à chaque mois.

Tableau 4
Taux moyen de divergence par question (définie au Tableau 4a)

Question	Taux moyen de divergence	Intervalle de variation des taux en 1984 (minimum-maximum)
10	0.2%	0.2% chaque mois
12	12.3%	10.4% à 14.3%
14	6.7%	5.7% à 8.4%
16	0.4%	0.3% à 0.5%
17	6.6%	2.0% à 9.9%
30	0.4%	0.3% à 0.5%
32	7.0%	3.0% à 11.6%
33	4.3%	1.8% à 6.0%
36	10.6%	8.1% à 12.7%
40	4.1%	1.5% à 6.8%
41	12.1%	6.2% à 19.7%
54	10.1%	7.9% à 12.1%
76	<0.1%	0.0% à 0.1%
77	15.0%	11.8% à 17.3%

Source: Rapport interne présenté à P.D. Changunde par Karen Switzer, 4 mars 1985, "Some Findings on the Field Edit Module (FEM) Reports from 1984".

Tableau 4a
Libellé des questions

- 10) La semaine dernière, (le répondant) a-t-il(elle) travaillé à un emploi ou à une entreprise peu importe le nombre d'heures? Oui ou Non.
- 12) Si la réponse à la question 11 "A-t-il(elle) plus d'un emploi la semaine dernière?" est oui, était-ce dû à un changement d'employeur? Oui ou Non.
- 14) Pourquoi ... travaille-t-il(elle) habituellement moins de 30 heures par semaine?, si la réponse à la question 13) indique un nombre d'heures travaillées inférieur à 30.
- 16) La semaine dernière, combien d'heures ... a-t-il(elle) été absent(e) du travail pour une raison quelconque (jour férié, vacances, maladie, conflit de travail, etc.)? (Ne pas oublier d'inscrire 00 s'il y a lieu).
- 17) Quelle était la raison principale de cette absence? (10 codes possibles)
- 30) La semaine dernière, ... avait-il(elle) un emploi ou une entreprise auquel il(elle) n'a pas travaillé? Oui ou Non.
- 32) A partir de la fin de semaine dernière, dans combien de semaines ... doit-il(elle) commencer à travailler à son nouvel emploi? (Si on a répondu Oui à la question 31), "La semaine, dernière, ... avait-il(elle) un emploi devant commencer à une date future déterminée?"
- 33) Pourquoi ... s'est-il(elle) absent(e) du travail la semaine dernière? (8 codes possibles)
- 36) Même que 14), sauf qu'elle concerne les personnes en chômage au lieu des personnes occupées. Au cours des 4 dernières semaines, ... s'est-il(elle) cherché(e) un autre emploi? Oui ou Non.
- 41) Qu'a fait ... au cours des quatre dernières semaines pour se trouver un autre emploi? (8 codes possibles, 1 à 3 codes différents dans 1, 2 ou 3 cases).
- 54) "Quelle est la raison principale pour laquelle ... a laissé cet emploi?" Neuf codes possibles si l'on a répondu oui à la question 50) " ... a-t-il(elle) déjà travaillé à un emploi ou à une entreprise?" (dans le cas des personnes souffrant d'incapacité permanente) et si l'on a répondu aux questions 51) à 53) qui concernent la date du dernier emploi et le statut d'employé (à temps plein ou à temps partiel). On omet la question 54) si la date du dernier emploi se situe avant la date pré-imprimée au n 52).
- 76/77) Indiquent, relativement à l'emploi principal (76) et à un autre emploi (77), quel est le statut du travailleur et si c'est le même qu'au mois précédent.

sur les questions et non sur les sous-questions, il est difficile de définir de façon catégorique le taux de divergence des réponses. Celui-ci a trait à un sous-ensemble de questionnaires pour lesquels une question particulière, disons la question $n^o q$, est pertinente selon des questions-filtres et des tables de décisions. Supposons que dans un échantillon de m questionnaires la question $n^o q$ soit pertinente pour $m_q \leq m$ questionnaires. Le taux de divergence est donc la proportion des m_q questionnaires qui ont été rejetés à la vérification à cause de la non-réponse à une question ou d'une erreur de codage. Il reste à savoir si le sous-ensemble m_q devrait inclure les questionnaires qui comptent des espaces remplis par erreur, ceux qui comptent des espaces laissés en blanc par erreur ou simplement ceux où les réponses sont codées correctement ou incorrectement. Malgré cet aspect ambigu de la définition, l'analyse des taux de divergence pour une cinquantaine de questions, pour l'année civile 1984, devrait indiquer une limite supérieure de l'erreur partielle dans les estimations des statistiques fondées sur les questions. Le Tableau 4 ci-dessous donne certains taux de divergence pour les questions définies au Tableau 4a, pour 1984.

Ainsi, pour une question aussi simple que la question 10 (le répondant a-t-il travaillé la semaine dernière, oui ou non), le taux de divergence n'est que de 0.2%, soit beaucoup moins que l'erreur type à l'échelle nationale. Pour des questions plus complexes comme 12, 36, 41, 54 et 77, le taux de divergence se situe en moyenne au-dessus de 10% et fluctue dans un intervalle de 2 à 6 points de part et d'autre de la moyenne au cours de l'année. On élimine les divergences par des méthodes du genre *hot deck*, l'utilisation de réponses fournies à la dernière enquête (si celles-ci sont disponibles) ou par la déduction logique appliquée à d'autres données du questionnaire. Par conséquent, il est souvent possible qu'une divergence soit interprétée comme le résultat d'une erreur de réponse et non d'une erreur d'imputation, de sorte qu'on devrait considérer les taux de divergence comme une limite supérieure de la fréquence globale des erreurs d'imputation dans les questions.

c) Non-réponse au questionnaire et correction de poids

En ce qui a trait à la non-réponse au questionnaire, les deux éléments de Y définis par (3.4) et (3.5) doivent être analysés ensemble puisque la non-réponse au questionnaire est normalement compensée par une correction de poids ((w_i)) plutôt que par une substitution directe de z_i à une valeur d'unité manquante. On calcule normalement les facteurs de correction de poids au moyen de l'inverse des taux s'appliquant à des cellules de correction; celles-ci se divisent en deux grandes catégories: les régions de compensation et les classes de pondération. Les premières sont souvent des régions géographiques liées au plan de sondage, comme une strate, une unité primaire d'échantillonnage, une grappe, un groupe de strates ou même un échantillon complet. Les classes de pondération sont définies par des strates formées à posteriori (c'est-à-dire, créées après l'échantillonnage) d'après des renseignements dont pouvaient disposer les répondants et les non-répondants compris dans l'échantillon. Pour ce qui a trait aux non-répondants, les renseignements peuvent être obtenus des unités qui ont répondu partiellement au questionnaire et dont on connaît certaines caractéristiques, peu importe que la caractéristique estimée se trouve ou non parmi celles-ci. On peut, par ailleurs, consulter des sources externes ayant trait aux non-répondants. On peut calculer l'inverse des taux de réponse soit pour des régions de compensation ou des classes de pondération et les utiliser comme facteurs de correction de poids pour compenser les cas de non-réponse dans les cellules.

Il existe plusieurs genres de facteurs de correction de poids pour compenser la non-réponse au questionnaire, le plus courant étant le taux de réponse inverse qui est défini comme le rapport de la taille de l'échantillon à la taille de l'échantillon répondant dans une cellule de

occupée ou inactive soit classée comme telle dans l'EPA est estimée à 0,9831 et à 0,9752 respectivement pour 1984, ce qui est légèrement inférieur aux probabilités correspondantes dans la CPS (0,9905 et 0,9923). On ne peut expliquer ces écarts pour le moment. De toute façon, les erreurs de réponse sont vraisemblablement plus importantes au niveau national qu'au niveau des petites régions. Par exemple, le biais dû à la réponse au niveau national peut être supérieur à l'erreur due à l'échantillonnage, tandis que l'estimation pour une petite région peut afficher un biais qui est comparable, en pourcentage, à celui obtenu à l'échelle nationale, mais beaucoup moins fort que l'erreur d'échantillonnage.

b) Non-réponse à une question et erreur d'imputation

Le troisième élément (3.3) de l'estimation X au Tableau 1 définit l'écart entre X et l'estimation voulue X qui est dû à l'imputation faite pour tenir compte de la non-réponse à une question, lorsque la valeur imputée $z_{ij} = y_i$ et que les écarts pondérés échantillonnés ($z_{ij} - y_i$) pour l'ensemble des unités échantillonnées qui font l'objet d'imputation pour la non-réponse à une question ne s'annulent pas. On considère qu'il y a non-réponse à une question lorsqu'une unité refuse de répondre à certaines questions du questionnaire ou que l'intervieweur ou le répondant (dans le cas de l'autodénonciation) laisse par mégarde des questions sans réponse. La deuxième cause de non-réponse à une question est identique à celle qui est à l'origine des erreurs de réponse, c'est-à-dire un questionnaire comportant des questions complexes et des définitions ambiguës ou un questionnaire confus où, suivant les réponses aux questions-filtres, on peut difficilement s'y retrouver.

Lorsqu'on a des cas de non-réponse à une question, on peut adopter une stratégie d'imputation comme celles qui ont été décrites précédemment; cette stratégie se traduit presque toujours par une substitution explicite. Dans l'analyse de microdonnées, il est essentiel d'obtenir une valeur z_{ij} qui se rapproche le plus possible de la valeur réelle Y_i ou, du moins, de ce qu'aurait été la valeur observée y_i si l'unité avait répondu à la question qui définit la caractéristique y . Il n'y a malheureusement aucun moyen de connaître le degré de concordance de z_{ij} et de y_i , si ce n'est par une réinterview de l'unité ou une analyse de sources externes ou de données d'enquêtes antérieures (qui peuvent être difficiles à obtenir). Un autre inconvénient de la non-réponse à une question et, par conséquent, de l'imputation est le faux sentiment de sécurité que celle-ci procure à l'utilisateur de données qui ignore peut-être qu'une valeur z_{ij} a été substituée à une réponse authentique au niveau des microdonnées. La valeur imputée z_{ij} tendra à s'écarter plus, dans un sens comme dans l'autre, de la valeur réelle Y_i , que l'erreur de réponse probable y_i si l'unité en question répond à la caractéristique. Il peut y avoir des exceptions. Malheureusement, il est normalement impossible de déterminer au niveau des microdonnées si z_{ij} constitue une valeur moins précise que y_i . Même s'il peut arriver que l'erreur d'imputation soit inférieure à l'erreur de réponse probable, cela réduira davantage la qualité des statistiques publiées à cause de la présence d'autres éléments de variance.

Les cas de non-réponse à une question et les erreurs de réponse dans l'EPA sont souvent répétés grâce au programme mensuel Module de contrôle sur le terrain qui permet d'analyser les questionnaires qui ont été rejetés à la vérification pour une ou plusieurs questions. En cours d'analyse, toutefois, on établit rarement la distinction entre les erreurs de réponse et les cas de non-réponse à une question, si l'on ne se donne pas la peine d'examiner à fond chaque questionnaire. L'erreur la plus courante est l'inscription d'un mauvais code plutôt que l'absence d'une réponse à une question. De nombreuses questions sont divisées en cinq ou six sous-questions, et il peut arriver qu'une erreur de codage soit considérée comme un cas de non-réponse à une sous-question et comme une erreur de réponse à une autre sous-question de la même question. Comme l'analyse du module du contrôle sur le terrain porte

Probabilité qu'une personne se soit déclarée occupée, en chômage ou inactif dans la CPS régulière, selon la situation réelle vis-à-vis de l'activité fondée sur les résultats de la réinterview de mai 1976.

	En chômage	Inactif
Total¹	0.9905	0.0016
Hommes²	0.9922	0.0013
Occupé	0.0474	0.8720
En chômage	0.0062	0.0048
Inactif	0.9890	0.0089
Femmes³	0.9892	0.0019
Occupé	0.0194	0.8442
En chômage	0.0049	0.0015
Inactif	0.1363	0.9936

³ Taille d'échantillon = 3 750

Source: Données établies à partir du "General Labour Force Status in the CPS Reinterview menée selon l'activité déclarée lors de l'interview originale.
Les deux sexes. Total. Après comparaison.
Mai 1976, Bureau of the Census (données non publiées).

Les deux sexes. Total. Après comparaison.

May 19/6, Bureau of the Census (données non publiées).

Nombre de personnes et probabilité de déclaration de l'activité
(entre parenthèses) selon la vraie caractéristique
(janvier-novembre 1984)

Caractéristique réelle (Réinterview de comparaison)	EPA régulière			
	Occupé	En chômage	Inactif	Total
Occupé	4 082 (0.9831)	19 (0.0046)	51 (0.0123)	4 152
En chômage	8 (0.0122)	571 (0.8691)	78 (0.1187)	657
Inactif	28 (0.0120)	30 (0.0128)	2 281 (0.9752)	2 339
Total	4 118	620	2 410	7 148

a) Erreur de réponse.

La somme des deux premiers éléments de l'estimation Y (voir 3.1 et 3.2) est égale à l'estimation d'Horvitz-Thompson qui sert à estimer le total lorsque le taux de non-réponse est nul. La réponse y_i fournie par l'unité peut différer de la valeur réelle X_i , ce qui entraîne une erreur de réponse pour l'unité i . À défaut de pouvoir l'éliminer, on peut diminuer l'erreur de réponse de deux façons seulement: par la formation adéquate des intervieweurs et par un questionnaire bien conçu, dénué de tout élément qui risquerait de déconcentrer l'intervieweur ou le répondant et où les caractéristiques sont clairement définies et les questions formulées en termes explicites.

Lorsque, pour un taux de non-réponse nul, les erreurs de réponse pondérées échantillon-nées de (3.2) ne s'annulent pas, l'estimation du total (Y) est entachée d'une erreur de réponse et, si l'on considère l'espérance mathématique E_1 et E_3 (voir Platek et Gray (1983)) pour tous les échantillons et les réponses possibles, on découvre que cette estimation comporte un biais dû à la réponse (BR) et une variance de réponse, sans compter la variance d'échantillonnage (VE). La variance de réponse peut être divisée en variance de réponse simple (VRS) et en variance de réponse corrélée (VRC).

Le biais dû à la réponse et tous les éléments de variance (VE, VRS et VRC) se rapportant à l'estimation ci-dessus sont calculés dans Platek et Gray (1983), paragraphe 2.2, p. 257-258. On analyse normalement les erreurs de réponse au moyen d'un programme de réinterview de comparaison, en vertu duquel on interviewe à nouveau un sous-échantillon des unités répondantes et on compare, pour la même période de référence de l'échantillon, les données obtenues avec celles de la première interview afin de déterminer laquelle des deux interviews produit les réponses les plus exactes. Les programmes de réinterview de comparaison sont prévus dans l'enquête sur la population active du Canada et la U.S. Current Population Survey (CPS), deux enquêtes mensuelles comparables visant à mesurer le chômage, l'emploi, etc. Par exemple, Poterba et Summers (1984) présentent au Tableau 2 les résultats d'une réinterview de comparaison menée en mai 1976 dans le cadre de la CPS; le sous-échantillon était composé de 3 329 hommes et de 3 750 femmes. En comparant les réponses fournies par ce sous-échantillon avec celles fournies à la première interview, on connaît la situation réelle d'un individu vis-à-vis de l'activité, de telle sorte qu'il est possible de savoir si cet individu a répondu correctement à la question lors de la première enquête de la CPS. Il est donc possible de connaître le nombre d'unités du sous-échantillon qui sont effectivement occupées et qui avaient été classées dans l'une ou l'autre des trois catégories (occupé, en chômage, inactif) dans l'enquête originale. À l'aide des trois valeurs obtenues, il est possible d'estimer, selon l'activité réelle, la proportion (ou la probabilité) de réponses justes et erronées (voir le tableau suivant).

Ainsi, d'après les résultats de la réinterview de comparaison, environ 87,20% des hommes effectivement en chômage ont été classés comme tels lors de la première enquête, tandis que 12,80% d'entre eux (soit 0,0474 + 0,0806) ont déclaré faussement être occupés ou inactifs. En conséquence, si y_i désigne la caractéristique en *chômage*, c'est-à-dire que $X_i = 1$ lorsque le répondant i est une personne de sexe masculin qui est effectivement en chômage, la probabilité que $y_i = 1$ est de 0,8720 alors que la probabilité que $y_i = 0$ est de 0,1280. Dans l'enquête sur la population active du Canada, l'échantillon visé par la réinterview de comparaison pour la période de janvier à novembre 1984 comprenait 7 148 personnes. Le Tableau 3 ci-dessous donne la probabilité que ces répondants aient déclaré être occupés, en chômage ou inactifs dans l'EPA régulière suivant leur activité réelle révélée par la réinterview. Ainsi, la probabilité qu'une personne effectivement sans emploi soit classée comme telle dans l'EPA est estimée à 0,8691, ce qui est sensiblement comparable à la probabilité correspondante dans la CPS (0,8602). De même, la probabilité qu'une personne effectivement

forme de la f.c.d. dans la population. En utilisant le facteur de correction de poids (wa_i), pour gonfler le poids t_i^{-1} de l'échantillon dans l'équation (2.4), on élimine l'effet d'amplification et, du même coup, on obtient une estimation différente mais plus réaliste de la f.c.d. Lorsqu'il n'y a aucun cas de non-réponse à une question et au questionnaire, les estimations (2.1) et (2.2) se réduisent à l'estimation d'Horvitz-Thompson (1952) qui est sans biais, exception faite des erreurs de réponse. Lorsqu'il manque des données et qu'on procède à une imputation, les estimations (2.1) et (2.2) risquent toutefois d'être biaisées pour des raisons autres que les erreurs de réponse, à moins que les valeurs de z_{ij} et de z_i se rapprochent de celles de y lorsqu'il faut procéder à une imputation pour la non-réponse à une question ou au questionnaire.

Dans la section suivante, les estimations (2.1) et (2.2) sont analysées en fonction de l'erreur de réponse, de l'erreur d'imputation due à la non-réponse à une question, de l'erreur d'imputation due à la non-réponse au questionnaire et de l'effet des facteurs de correction de poids supérieurs à 1.

3. ÉLÉMENTS DE L'ESTIMATION

L'estimation \bar{Y} définie en (2.1) ou (2.2) peut être décomposée en 5 éléments, dont le premier est l'estimation d'Horvitz-Thompson, laquelle, illustrée au Tableau 1, utilise les valeurs réelles de la caractéristique. La f.c.d. estimée $F(Y)$, définie en (2.4), peut être décomposée de la même façon; cependant, nous n'avons pas jugé bon de traiter ce sujet dans le présent document.

Lorsque le facteur de correction de poids ($wa_i = 1$, le dernier élément disparaît et la somme des 4 autres ((3.1) à (3.4)), correspond à l'estimation définie en (2.1). Lorsque la non-réponse au questionnaire est compensée par un facteur de correction de poids supérieur à 1, la valeur z_i n'est pas imputée directement pour la valeur manquante et elle est considérée comme nulle dans l'élément (3.4). Dans ce cas, la somme des 5 éléments correspond à l'estimation définie en (2.2) et l'effet négatif de la non-réponse au questionnaire en (3.4) est compensé par l'effet positif de la correction de poids en (3.5).

Tableau 1:
Éléments de l'estimation \bar{Y}

(3.1)	Estimation sans biais fondée sur un taux de réponse de 100% et des valeurs réelles	..	$\bar{Y} = \sum_{i=1}^I t_i \pi_i^{-1} Y_i$	
(3.2)	effet de l'erreur de réponse	..	$+ \sum_{i=1}^I t_i \pi_i^{-1} (Y_i - Y_j)$	
(3.3)	effet de la non-réponse à une question	..	$+ \sum_{i=1}^I t_i \pi_i^{-1} \delta_i (1 - \delta_{ij})(z_{ij} - Y_j)$	
(3.4)	effet de la non-réponse au questionnaire	..	$+ \sum_{i=1}^I t_i \pi_i^{-1} (1 - \delta_i)(z_i - Y_i)$	
(3.5)	effet de la correction de poids pour la non- réponse au questionnaire	..	$+ \sum_{i=1}^I t_i \pi_i^{-1} [wa_i - 1] \delta_i [\delta_{ij} Y_j + (1 - \delta_{ij}) z_{ij}]$	

L'aide d'une table de décision. On peut aussi obtenir une valeur imputée à partir de données fournies antérieurement par la même unité au cours d'une enquête ou d'un recensement ou à partir de données administratives, si de telles données existent. On peut appliquer des méthodes de régression, la déduction logique et bien d'autres encore. Des erreurs systématiques sont parfois attribuables à de mauvaises opérations de préposés au codage ou à la perforation. Dans ce cas, on peut vouloir remplacer les codes par des valeurs logiques se rapportant à d'autres données contenues dans le questionnaire, au lieu de procéder à une imputation. Quoi qu'il en soit, le but ultime est d'obtenir une valeur imputée ou un code modifié qui se rapproche le plus de la valeur réelle Y_i . Dans les enquêtes permanentes, où les caractéristiques étudiées évoluent peu sur une longue période (par exemple, l'emploi dans certaines industries et certaines professions), on peut considérer les réponses fournies aux enquêtes antérieures par un répondant comme presque aussi valables que celles fournies au cours de l'enquête courante, à plus forte raison lorsque les périodes de référence des deux enquêtes ne sont pas trop éloignées. Il peut en être de même pour les données d'enquête recueillies à un an d'intervalle, comme celles ayant trait à des caractéristiques saisonnières (par exemple, l'industrie de la pêche). Dans les cas de non-réponse au questionnaire, on peut parfois imputer des données d'enquêtes antérieures si le non-répondant en question a participé à ces enquêtes et qu'il présente des caractéristiques stables.

Pour les cas de non-réponse au questionnaire, l'imputation se fait généralement par une correction de poids au moyen de l'inverse du taux de réponse enregistré dans une cellule ou une région. L'estimation du total est donc définie comme suit:

$$Y = \sum_{i=1}^I t_i \pi_i^{-1} (w_i) \delta_i [\delta_{iy} Y_i + (1 - \delta_{iy}) z_{iy}] \quad (2.2)$$

où (w_i) est le facteur de correction du poids pour l'unité i , qui vise à rééquilibrer l'échantillon amené par la non-réponse au questionnaire. Dans l'expression ci-dessus, on suppose que toutes les questions laissées sans réponse par l'unité répondante i ($\delta_{iy} = 0$) ont déjà fait l'objet d'une imputation au moyen de z_{iy} .

En supposant qu'il puisse manquer des données, on peut obtenir des estimations de la fonction cumulative de distribution (f.c.d.) dans l'échantillon en remplaçant la valeur observée y par la variable auxiliaire $c(y, Y)$, celle-ci étant égale à 1 ou à 0 selon que y est inférieure ou supérieure à Y ; on fait de même pour z_{iy} et z_{ij} . Les f.c.d. correspondant à (2.1) et (2.2) sont estimées respectivement par les équations (2.3) et (2.4).

$$F(Y) = \frac{1}{N} \sum_{i=1}^I t_i \pi_i^{-1} \{ \delta_i [\delta_{iy} c(y, Y) + (1 - \delta_{iy}) c(z_{iy}, Y)] + (1 - \delta_i) c(z_{iy}, Y) \} \quad (2.3)$$

où $N = \sum_{i=1}^I t_i \pi_i^{-1}$ représente le nombre réel ou estimé d'unités dans l'univers. Ainsi, suivant la base de sondage, le plan de sondage et le listage des unités, N peut ou non évaluer N .

$$F(Y) = \frac{1}{N} \sum_{i=1}^I t_i \pi_i^{-1} (w_i) \delta_i [\delta_{iy} c(y, Y) + (1 - \delta_{iy}) c(z_{iy}, Y)] \quad (2.4)$$

Tandis que les valeurs de Y définies en (2.1) et (2.2) sont identiques, peu importe que l'imputation pour la non-réponse au questionnaire se fasse par l'application de moyennes relatives aux répondants ou par la correction de poids, les estimations $F(Y)$ des f.c.d., définies en (2.3) et (2.4) ne le sont pas. Lorsqu'on impute en appliquant les moyennes relatives aux répondants (moyenne globale ou moyennes s'appliquant à des cellules de correction définies pour la compensation de la non-réponse), comme dans l'équation (2.1) ou (2.3), on se trouve à amplifier ces valeurs moyennes dans la f.c.d. estimée, ce qui a pour effet de fausser la

La non-réponse à une question est un problème souvent plus difficile à traiter que la non-réponse au questionnaire, sur laquelle a porté surtout notre propos jusqu'à maintenant. Les deux principaux facteurs qui peuvent réduire le taux de non-réponse à une question sont un questionnaire bien conçu et des intervieweurs très compétents, c'est-à-dire recrutés et formés selon les critères appropriés. Un questionnaire mal conçu peut amener un intervieweur ou un répondant à s'interroger sur la façon exacte de remplir ce questionnaire. La non-réponse à une question est donc possible malgré la bonne volonté de l'intervieweur ou du répondant. En outre, il se peut que des répondants consentent à répondre à certaines questions seulement. Quelle que soit la cause des réponses manquantes, le problème de leur substitution reste entier. En règle générale, un organisme d'enquête est peu disposé à rejeter les renseignements qu'il vient de recueillir à moins, bien sûr, que les réponses fournies aux principales questions ne semblent tout à fait erronées ou illogiques. On a donc recours habituellement à des méthodes d'imputation pour les réponses manquantes tout en conservant les réponses déjà obtenues.

Une enquête ou un recensement doivent produire diverses statistiques permettant d'expliquer des phénomènes sociaux, de déterminer des programmes socio-économiques, etc. Ces statistiques comprennent des moyennes, des totaux, des ratios, des distributions, des percentiles et des graphiques. On suppose que ces statistiques sont fondées sur un univers de N unités appartenant à la population visée, où N peut être connu ou non. Il est possible de démontrer que toutes ces statistiques peuvent être exprimées en fonction de totaux ou de comptages. Ainsi, le reste de cet article est consacré à l'analyse des données manquantes dans la mesure où celles-ci influent sur les estimations des totaux et des comptages dans les enquêtes. Certaines observations porteront aussi sur les recensements.

2. FORMULE D'ESTIMATION

Lorsqu'il y a non-réponse à une question ou au questionnaire, le total de la caractéristique y peut être estimé au moyen de l'expression générale suivante:

$$\bar{Y} = \sum_{i=1}^N t_i \pi_i^{-1} \{ \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}] + (1 - \delta_i) z_i \}, \text{ où} \quad (2.1)$$

t_i = 1 ou 0, selon que l'unité i est choisie ou non;
 π_i = probabilité de sélection de l'unité;
 δ_i = 1 ou 0, selon que l'unité i répond ou non;
 δ_{iy} = 1 ou 0, selon que l'unité répondante i répond ou non à la question ou à la caractéristique y ;

y_i = réponse observée pour la caractéristique y lorsque $\delta_{iy} = \delta_i = 1$;
 y_i peut être ou non égal à Y_i , la valeur réelle;
 z_{iy} = valeur imputée pour la non-réponse à une question, lorsque

$\delta_i = 1, \delta_{iy} = 0$.
 z_i = valeur imputée pour la non-réponse au questionnaire, lorsque

$\delta_i = 0$.

L'estimation ci-dessus peut se rapporter à une catégorie a d'unités; dans ce cas, la variable auxiliaire β_{ia} insérée après π_i^{-1} dans l'équation, prend la valeur 1 ou la valeur 0 selon que l'unité i appartient ou non à la catégorie a (par exemple, catégorie d'âge selon le sexe). Pour ce qui a trait à la non-réponse à une question, z_{iy} est presque toujours une valeur imputée explicite pour les données manquantes. On peut obtenir une valeur imputée par la méthode *hot deck*, c'est-à-dire qu'on applique à une question y , laissée sans réponse par l'unité i , une réponse fournie par une autre unité qui ressemble le plus possible à l'unité i , à

Méthodes de compensation de la non-réponse

R. PLATEK et G.B. GRAY¹

RÉSUMÉ

La non-réponse à une question ou à un questionnaire est un phénomène presque inévitable dans les enquêtes et les recensements. Plus elle a de l'ampleur, plus elle risque d'influer sur les estimations de l'enquête. Il importe donc de contenir ses effets à toutes les étapes où elle peut fausser les estimations. À des degrés divers, il est possible de minimiser l'ampleur de la non-réponse aux stades de la conception, de l'exécution et du traitement. Les problèmes de non-réponse se reflètent sur les formules d'estimation relatives à diverses statistiques à cause des imputations et des corrections de poids, sans compter les poids attribués dans une enquête aux estimations de moyennes, de totaux ou d'autres statistiques. Les formules peuvent être décomposées en divers éléments comme les erreurs de réponse, l'effet de la correction de poids pour tenir compte de la non-réponse au questionnaire et l'effet de la substitution pour non-réponse. Les auteurs analysent l'effet des diverses étapes (conception, exécution et traitement) sur les éléments des estimations.

MOTS CLÉS: Non-réponse; imputation; estimation.

1. INTRODUCTION

Dans la collecte de données auprès d'unités échantillonnées, il y a toujours un certain nombre d'unités qui négligent de répondre à une question ou au questionnaire malgré tous les efforts faits pour que cela ne se produise pas. Le problème que pose le traitement de la non-réponse et des données manquantes qui en découlent est double. Tout d'abord, il faut déterminer dans quelle mesure on peut réduire l'erreur quadratique moyenne des données d'enquête par des visites de rappel, des lettres de rappel, etc. sans que cela n'accroisse indûment le coût de l'enquête et, ensuite, il faut déterminer le genre d'ajustements requis pour les données manquantes découlant des cas persistants de non-réponse, afin de réduire le biais dû à la non-réponse. Pour réduire au minimum le taux de non-réponse au questionnaire, les bureaux régionaux ou l'administration centrale des enquêtes doivent tenter de contacter à plusieurs reprises les unités choisies jusqu'à ce qu'une personne responsable puisse répondre au questionnaire d'enquête. De telles mesures sont prises dans le cas des interviews sur place ou des interviews téléphoniques. En ce qui concerne les enquêtes par la poste, ce genre de démarche se traduit par l'expédition répétée du questionnaire d'enquête aux unités non répondantes. Cela peut parfois donner lieu à un suivi téléphonique ou à un suivi à domicile. Bien qu'il subsistara toujours des cas de non-réponse, on devrait faire tous les efforts possibles pour réduire leur nombre au minimum. Ainsi, il y aura toujours des non-répondants qui résisteront aux demandes répétées des intervieweurs, peu importe les mesures que ceux-ci prendront. Il est donc nécessaire d'appliquer une méthode d'imputation pour tenir compte des données manquantes. Dans le présent document, nous nous penchons sur les problèmes que soulève le contrôle de la non-réponse à l'étape de la conception et de l'exécution d'une enquête, puis nous analysons certaines méthodes de compensation de la non-réponse à l'étape du traitement. Dans cette deuxième partie, nous examinons l'applicabilité de ces méthodes de même que les questions d'ordre pratique et méthodologique qui s'y rattachent.

¹ R. Platek et G.B. Gray, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, 4^e étage, Imm. Jean-Talon, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 11, numéro 1, juin 1985

TABLE DES MATIÈRES

R. PLATEK et G.B. GRAY	1
Méthodes de compensation de la non-réponse.....	1
J.N.K. RAO	17
Inférence conditionnelle dans les enquêtes par sondage.....	17
G.H. CHOUDHRY, H. LEE, et J.D. DREW	37
Optimisation du coût et de la variance dans le cadre de l'enquête sur la population active au Canada.....	37
K. CHIU, J. HIGGINSON, et G. HUOT	57
Évaluation de modèles ARMMI appliqués à des séries chronologiques.....	57
M.A. HIDIROGLOU et C.E. SÄRNDAAL	73
Étude empirique de quelques estimateurs de régression pour petits domaines.....	73
D.K. HOLINS	87
Méthode de traitement des données du recensement de l'agriculture de 1981.....	87

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

COMITÉ DE RÉDACTION

Président	R. Platek	Statistique Canada
Rédacteur en chef	M.P. Singh	Statistique Canada
Rédacteurs associés	K.G. Basavarajappa	Statistique Canada
	D.R. Bellhouse	Université Western Ontario
	E.B. Dagum	Statistique Canada
	J.F. Gentleman	Statistique Canada
	G.J.C. Hole	Statistique Canada
	T.M. Jeays	Statistique Canada
	G. Kalton	Université du Michigan
	C. Patrick	Statistique Canada
	J.N.K. Rao	Université Carleton
	C.E. Sarnadal	Université de Montreal
	V. Tremblay	Université de Montreal
Rédacteur adjoint	H. Lee	Statistique Canada

COMITÉ DE DIRECTION

R. Platek (Président), E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques-qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les techniques de lis-sage et d'extrapolation, les études démographiques, l'intégration et l'analyse de production de statistiques. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles sont soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, 4^e étage, Edifice Jean-Talon, Parc Tunney, Ottawa (Ontario), Canada KIA 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Statistique Canada

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

Juin 1985

Publication autorisée par
le ministre des Approvisionnements
et Services Canada

© Ministère des Approvisionnements
et Services Canada 1985

Décembre 1985
8-3200-501

Prix: Canada, \$10.00, \$20.00 par année
Autres pays, \$11.50, \$23.00 par année

Paiement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 11, n° 1

ISSN 0714-0045

Ottawa

Canada

VOLUME 11, NUMÉRO 1
JUN 1985

UNE REVUE
DE
STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



2-00/

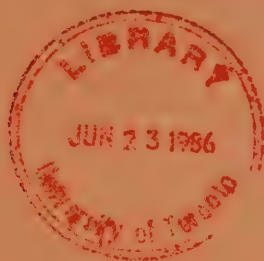


Statistics Canada Statistique Canada

Government
Publications

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA



VOLUME 11, NUMBER 2
DECEMBER 1985

Canada

Statistics Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

December 1985

Published under the authority of
the Minister of Supply and
Services Canada

© Minister of Supply
and Services Canada 1986

May 1986
8-3200-501

Price: Canada, \$10.00, \$20.00 a year
Other Countries, \$11.50, \$23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 11, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

EDITORIAL BOARD

Chairman	R. Platek, <i>Statistics Canada</i>
Editor	M.P. Singh, <i>Statistics Canada</i>
Associate Editors	K.G. Basavarajappa, <i>Statistics Canada</i> D.R. Bellhouse, <i>University of Western Ontario</i> E.B. Dagum, <i>Statistics Canada</i> J.F. Gentleman, <i>Statistics Canada</i> G.J.C. Hole, <i>Statistics Canada</i> T.M. Jeays, <i>Statistics Canada</i> G. Kalton, <i>University of Michigan</i> C. Patrick, <i>Statistics Canada</i> J.N.K. Rao, <i>Carleton University</i> C.E. Särndal, <i>University of Montreal</i> V. Tremblay, <i>University of Montreal</i>
Assistant Editor	H. Lee, <i>Statistics Canada</i>

MANAGEMENT BOARD

R. Platek (Chairman), E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

EDITORIAL POLICY

The Survey Methodology Journal will publish articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, smoothing and extrapolation methods, demographic studies, data integration and analysis and related computer systems development and applications. The emphasis will be on the development and evaluation of specific methodologies as applied to actual data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$10.00 per copy, \$20.00 per year in Canada, \$11.50 per copy, \$23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales and Services, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 11, Number 2, December 1985

CONTENTS

M.B. WILK	
The Relationship between Statisticians and Statisticians	89
J.D. DREW, Y. BÉLANGER, and P. FOY	
Stratification in the Canadian Labour Force Survey	95
D.R. BELLHOUSE	
Sampling Microfilmed Manuscript Census Returns	111
B.C. SAXENA, P. NARAIN, and A.K. SRIVASTAVA	
Estimation of Total for Two Characters in Multiple Frame Surveys	119
E.B. DAGUM and M. MORRY	
Seasonal Adjustment of Labour Force Series during Recession and Non-Recession Periods	133
E.B. DAGUM, G. HUOT, N. GAIT, and N. LANIEL	
Relational Patterns between Total Unemployment and Unemployment Beneficiaries in Canada	145
L. SWAIN	
Basic Principles of Questionnaire Design	161
R.B.P. VERMA and P. PARENT	
An Overview of the Strengths and Weaknesses of the Selected Administrative Data Files	171
R.D. SHARMA and C. WONG	
Use of Administrative Data Files for Migration Estimates; A Case Study of Driver's Licence File in Ontario	181
F. AHMAD, R. CHOW, O. DEVRIES, A. HASHMI, and Y. MARCOGLIESE	
The Development of Alberta Health Care Records and Their Applications to Small-Area Population Estimates	187
D.G. McRAE	
The Use of Hydro Accounts in the British Columbia Regression Based Population Estimation Model	197
D.S. O'NEIL and C.D. McINTOSH	
Estimating the Age/Sex Distribution of Small Area Population	203
R.B.P. VERMA, K.G. BASAVARAJAPPA, and R.K. BENDER	
Estimating Population by Age and Sex for Census Divisions and Census Metropolitan Areas	211
R.K. BENDER	
Experience with Small Area Population Estimates	219
Editorial Collaborators	223

The Relationship between Statisticians and Statistics¹

MARTIN B. WILK²

I appreciate the honour of the invitation as after-dinner speaker at this 1985 annual meeting of the Statistical Society of Canada.

The honour is unfortunately accompanied by a responsibility, to say something worthwhile. That is not an easy task. I thought I would approach that job in stages. So first I invented a title. Then I thought I would try to figure out what the title meant. And that was to be my speech. Regrettably, I am still unsure what the title means. But I won't let that deter me. Of course, as Yogi Berra said, "If you don't know where you're going you may not get there".

There are many people called statisticians who carry out a very diverse set of activities which are labelled statistics. In fact, at various times in my unplanned career, I have been various kinds of statisticians. That fact of language poses the question: What are the relationships among these various kinds of statisticians and statistics?

Specifically let me identify two types of statistical activity, namely probability statistics on the one hand and the work of statistical information development, carried out by statistical agencies, on the other hand. What do I mean by probabilistic statistics? Without any attempt to be precise, I mean to encompass the discipline commonly covered in standard texts and lectures including notions of analyses of variance, tests of goodness of fit, design of experiments, variance components, Bayesian estimation and so forth.

The results of the work of statistical agencies, like Statistics Canada and the Manitoba Bureau of Statistics and the U.S. Bureau of the Census, you read in the newspapers every day.

These two kinds of work are *perceived* as related, and I believe *are* related. You might say the relationship has both a *real* and an *imaginary* part – and I am not at all clear what aspects fall into which category.

Let us take a look at some of the manifestations of these two categories – which one might also label as *white collar statistics* and *blue collar statistics* (which terms are used purely to avoid laborious repetition of awkward phrases like "probability statisticians").

The Statistical Society of Canada seems to be predominantly an organization of white collar statisticians. A recent study indicated

66% academic membership

21% government agencies.

The Statistical Society of Canada lists 32 persons from Statistics Canada as members, out of 2,000 professionals.

¹ Invited address at the annual meeting of the Statistical Society of Canada, Winnipeg, Manitoba, June 1985.

² Martin B. Wilk, formerly Chief Statistician of Canada, Currently Senior Advisor to Privy Council and President of the Statistical Society of Canada.

Registration at this meeting likely consists mainly of white collar statisticians – interested primarily in the arena of probability statistics. Not only are there only a very few persons (8) from Statistics Canada, I must also report that there was only minimal interest of supervisors at Statistics Canada in sending persons to the meeting.

Let us look at examples of output from these two categories. The official journal of the Statistical Society of Canada is the Canadian Journal of Statistics. It is a quarterly. The official release announcement vehicle of Statistics Canada is the daily, which appeared 256 times last year.

A comparison of titles of publications is fascinating. For the Canadian Journal of Statistics, I selected at random fifteen key words from 122 which represented the articles published in 1983.

Here is a sample list of what white collar statisticians are writing and reading about:

- Abundance distributions
- Asymptotic properties
- Central Wishart distribution
- Chi-squared distribution
- Critical values
- Decision theory
- Growth-curve analysis
- Linear filter
- Logistic process
- Longitudinal studies
- Multivariate linear model
- Shift estimation
- Spatial time series
- Structural properties
- Weighted least-squares estimator

Those topics are household words at this conference. But they are *not* the topics of blue collar statistical output – and many, perhaps most, blue collar statisticians would have no understanding of, or concern with, these topics, at all.

Some indication of the output of Statistics Canada is provided by the releases announced in the daily of April 29, 1985.

- total number of pigs in Canada (over 10 million)
- the number of tonnes of barley exported (over 150,000 during March 1985)
- the number of square metres of mineral wool shipped (over 6 million)

A further indication of Statistics Canada output is the table of major statistical indicators, which is updated each week in a publication, statistical highlights, sent to ministers and deputy ministers. These indicators include:

- Gross National Product
- Housing Starts
- Bank Rate
- Unemployment Rate
- Consumer Price Index Increase
- Weekly Earnings

And the measures relating to economic, business, trade, financial, social and labour sectors of Canadian Society.

Statistics Canada turns out statistical studies on topics such as divorce in Canada, health of Canadians, the status of women, current economic indicators, science and technology indicators, language characteristics of Canadians and so on.

I want to make it clear that I am *not* engaged in making an assessment of the relative value of these two types of outputs. Both types of work are socially desirable, as indicated by the fact each has supporting social constituencies. By definition, each is socially justified.

But what I *am* engaged in is trying to analyze the nature of relationship between these two types of activities, both of which are labelled *statistics* and carried out by people who are called *statisticians*.

We could of course simply write it off as a case of homonymism – that is the same word being used with two entirely different meanings. Or we should simply continue to ignore this discrepancy. But neither of those is wise or productive.

You are all familiar with the classic work on the advanced theory of statistics by Kendall and Stewart. Volume I involves 396 pages of text plus tables and index. These 396 pages deal with theoretical constructs of probability statistics and mathematical derivations of various formulae.

The introductory quotation to the book is attributed to O. Henry and reads as follows:

“Let us sit on this log at the roadside”, says I, “and forget the inhumanity and ribaldry of the poets. It is in the glorious columns of ascertained facts and legalized measures that beauty is to be found. In this very log we sit upon, Mrs. Sampson,” says I, “is statistics more wonderful than any poem. The rings show it was sixty year old. At the depth of two thousand feet it would become coal in three thousand years. The deepest coal mine in the world is at Killingworth, near Newcastle. A box four feet long, three feet wide, and two feet eight inches deep will hold one ton of coal. If an artery is cut, compress it above the wound. A man’s leg contains thirty bones. The tower of London was burned in 1841.”

“Go on, Mr. Pratt”, says Mrs. Sampson. “Them ideas is so original and soothing. I think statistics are just as lovely as they can be.” (The handbook of Hymen).

I think the quotation is lovely. And the book is, of course, an excellent example of scholarly clarity. But I do wonder what is the connection between the quotation and the text? Do the authors see a close connection? Is the quotation – which reflects work like that of the blue collar statistician-intended to justify, or motivate, the superstructure of probabilistic statistics which follows?

Do the authors believe that the constructs and formulae of their text on probabilistic statistics serve to guide or validate the work of blue collar statisticians – of statistical agencies? Or do they believe that the discipline of probability statistics is justified because its technology has been used to produce the output of statistical agencies?

What is *real* and what is *imaginary* in this relationship?

There is something of a conundrum in the relationships between the work of white collar statistics and blue collar statistics. The apparent outlook seems to be that:

- The information product is valid because it uses approved methodology.
- The methodology has status because it derives from a formulated theory.
- But the statistical theory involves constructs and mathematical logic, usually based on various unverifiable assumptions.!

What justifies the assumptions, the constructs and the theory?

In scientific work, more generally, a theory is justified as good by the usefulness of the products produced by technology derived from the theory.

Indeed, technology is often invented without *theory* and widely accepted because of its utility. Bronze and Damascus steel were developed because of their useful properties, and not because of a mathematically consistent theory of metallurgy.

To assess whether probabilistic statistics is good, we should ask whether it provides a technology to produce products that are useful and valuable.

Instead, statisticians tend to ask the inverse question, namely whether the work of blue collar statistics is valid according to the precepts of probability statistics.

Probability statistics has produced a wide variety of concepts and models and methodologies. These include areas such as:

- Decision making under uncertainty
- Subjective probability
- Science of inference
- Likelihood inference
- Bayesian estimation
- Time series analysis
- Hypothesis testing
- Tests of significance
- Confidence estimation
- Estimation of sampling errors
- Classification methods
- Regression analysis
- Variance components
- Design of experiments
- Sample survey design
- Unbiased estimators

and so on.

Many authors have asserted that the most fundamental concept in applied probabilistic statistics is the *objective assessment of uncertainty*.

But I must tell you that that notion – however appealing and philosophically profound – does not comport with the reality of the work and mandate of statistical agencies.

Let me try to establish by example the social importance of the work of blue collar statisticians. You can make a test of your own. Make a list of what you believe to be the issues of interest to Canadian Society. Your list will include matters of employment and unemployment, income of the elderly, status of women, economic growth, trade and balance of payments, family formation, population distribution, government deficit, etc.

On examination you will find that, for the large majority of such issues, your perceptions, your knowledge and your understandings depend quite directly on the statistical information produced by blue collar statisticians, *mainly* at Statistics Canada. A similar assessment would apply in any country in the world.

To emphasize this point further by a specific example, I would like to summarize some of the uses of the consumer price index.

The consumer price index is updated each month by Statistics Canada based on monthly observations of prices of a designated market basket of goods and services. The consumer price index is the most commonly used indicator of the rate of inflation. It is often referred to as the cost of living index. The consumer price index has a direct or indirect effect on nearly all Canadians. It, or individual components of which it is weighted average, is used in the calculations or definitions of income taxes, labour contracts, family allowance payments, old age security pensions, rental agreements, insurance coverage, spousal support payments, child support payments, payments to children of war veterans, student loan repayments, and many other contractual or regulatory arrangements.

To get back to the matter of objective error estimation – supposedly the central feature of probability statistics: Statistics Canada does not produce a statistical measure of the error of the consumer price index estimate. We do *not* publish interval estimates of consumer price index. We do not test the hypothesis of no change in consumer price index from month to month. We do not produce composite estimates which would supposedly reduce random error variance.

From time to time we are queried or criticized about this, even by people who are not statisticians or scientists. It seems that, having heard so often about the results of public opinion polls, members of the public have now begun to expect an error estimate to accompany published estimates. The phrase "19 times out of 20" is now a part of the vocabulary of most newspaper readers. Of course, public opinion polls have been going on for a long time; George Gallup found a record of one taken back in 1824, when a pennsylvania newspaper published results of what was called a "straw vote taken without discrimination of parties". Modern communications and computer technology have resulted in a proliferation of polls. Because of their popularity, there has been an increase in public awareness of the fact that a statistician (or somebody.) can conduct a sample survey, make inferences, and put a measure of uncertainty on estimates.

An audit of Statistics Canada in 1983 by the Auditor General of Canada, touched on the subject of measuring the quality of statistics. The report recommended that Statistics Canada develop and disclose more measures of quality for its statistics. The agency's formal reply was that this "recommendation could not be fully implemented, since 'measures of quality' for many statistics - particularly those of a composite nature - are impossible to produce". It would be more realistic, said the Statistics Canada response, to supply "a full *description* of available information related to possible quality limitations, including, of course, quality measures when they are available".

Statistics Canada would publish more error estimates if we felt we could. It is not that we would mind admitting the possibility of error. As professor R. C. Bose used to say to his students "to err is human. Therefore, statisticians are human".

However, the usual error estimates depend on assumptions which vastly oversimplify the situation. For example, the labour force sample households, not independent individuals who have equal chances of being selected. Also, by design, the households themselves do not have equal chances of being sampled; the sampling ratio is approximately 1 in 125 at the national level, but can be as high as 1 in 24 for provinces with small populations. Can we assume, then, that all individuals are independent and have an equal probability of being unemployed? Data are gathered by means of an interview, and either the interviewer or the respondent may make an inadvertent or even a deliberate mistake. Can we ignore all possible sources of error except sampling error? Members of a given household are sampled for six consecutive months, with 1.6 of the households rotating into and out of the overall sample each month. Thus, in any month, different respondents have responded to the questionnaire different numbers of times. Can we assume that the six responses are independent over time? Sometimes, during the six months of sampling, families move away from, or move into, a particular dwelling being sampled. And, of course, there are the usual problems of non-response, outliers, and errors of data entry, computation, and printing, etc. Concern about how to handle deviations from the "usual" assumption of statistical theory is a major continuing preoccupation of some of Statistics Canada's blue and white collar statisticians.

So, on the one hand, probability statistics has contributed the appealing and important concept of objective estimation of error; and moreover the public has been educated to accept the concept and to expect it to be implemented.

On the other hand, there are many very influential and prominent statistical products produced by people called statisticians for which such measures *are not* provided, and cannot be provided at the present time.

Abstractly, there seem to be several options!

- (a) Statistical agencies and probability statistics might agree to stop sharing the label "statistics" and abandon the notion of connectivity.

- (b) Probability statistics could address its efforts to produce technology to deal with the reality of complex statistical information development.
- (c) Statisticians might undertake a public reeducation campaign to cancel the beliefs that neat and objective measures of statistical uncertainty are possible.

As a practical matter, only option 2 can be considered. And it also holds greater promise of productive consequences for *all* statisticians of all varieties.

In an article in science last year, Ian Hacking, a philosopher of science, commented that "the quiet statisticians have changed our world – not by discovering new facts or technical developments but by changing the ways we reason, experiment and form our opinions about it".

It is gratifying to read such an assessment of the significance of probabilistic statistics as pioneered by Fisher, Neyman, Pearson, Wald and others.

But, in the vein of my topic tonight, I want to point out that there is another cadre of "quiet statisticians" – the blue collar statisticians of statistical agencies – who have also contributed to changing the world; but precisely in the manner inverse to Mr. Hacking's assessment.

Blue collar statisticians *do* discover new facts.

They *do* establish new concepts.

They *do* invent operational definitions and implement them for public consumption.

They *do* pioneer technical developments – in computing, electronic dissemination of information, computer graphics, classification systems, national accounting frameworks and so on.

Again I want to remind you my intention is not to make, or to imply, an assessment of comparative value. The issue is: what is *real* and what is *imaginary* in the relationship of blue collar statistics and white collar statistics?

Most of what blue collar statisticians do does not in reality derive from, or directly relate to, the constructs and theories and beliefs associated with probability statistics. And yet – the blue collar statisticians are somehow persuaded or coerced into paying lip service to a supposedly fundamental connectivity to those concepts.

At the same time, the white collar statisticians continue with a vague belief that if only more of the blue collar statisticians could achieve academic respectability then probability statistics would *really* impact importantly on statistical agencies.

The synergy which may be latent in the more effective relationship of the blue collar and white collar statisticians will not be developed without effort from both groups.

I don't have the wisdom to offer any revelatory proposals.

Better channels of communication are obviously needed. In that spirit, Statistics Canada has established a program of fellowships and internships.

Also in that spirit, Statistics Canada has established a network of advisory committees, including one on statistical methodology.

A number of probabilistic statisticians are on contract as consultants to Statistics Canada.

I expect there is much more opportunity for expanding seminar exchanges and working collaborations between Statistics Canada and Universities.

There is a need for improved intellectual tolerance in both groups. Perhaps the criteria and standards for publishing need to be modified.

Perhaps the basis for judging the acceptability of research grants by the Natural Sciences and Engineering Research Council of Canada should be changed.

Perhaps training programs could usefully be modified. Perhaps Statistics Canada should offer a prize for productive developments related to outstanding areas of need in the operation of statistical agencies. Maybe we should have a continuing list of the ten most wanted solutions as an incentive, and communication mode, to probability statistics researchers.

Maybe the Statistical Society of Canada should establish a tradition that every year the after-dinner speaker at the annual meeting should talk about "the relationship of statisticians and statisticians".

Stratification in the Canadian Labour Force Survey

J.D. DREW, Y. BÉLANGER and P. FOY¹

ABSTRACT

The use of a multivariate clustering algorithm to perform stratification for the Labour Force Survey is described. The algorithm developed by Friedman and Rubin (1967) is modified to allow the formation of geographically contiguous strata and to delineate heterogeneous but compact primary sampling units (PSUs) within these strata. Studies dealing with stratification variables, stratification robustness over time, and type of stratification are described.

KEY WORDS: Multivariate clustering algorithm; Geographic stratification; Continuous survey.

1. INTRODUCTION

The Canadian Labour Force Survey is redesigned after every decennial census of population and housing. The redesign which occurred following the 1981 Census included an intensive program of research on various aspects of the sample design (Singh, Drew and Choudhry 1984). This report describes the portion of the research program dealing with stratification methods.

Because the LFS is used not only to provide information on labour force characteristics but also as a general design for various other household surveys, one of the principal objectives of the redesign was to increase the flexibility of the LFS for general applications. Stratification was considered a means of improving efficiency for general applications, as well as variables of particular interest to the LFS, through the application of more rigorous procedures than those used in the old design.

It was therefore decided to consider the use of multivariate clustering algorithms and to compare them with the methods used in the old design. A non-hierarchical algorithm developed by Friedman and Rubin (1967) was selected on the basis of the results of evaluations of various algorithms by Judkins and Singh (1981) as part of the redesign of the Current Population Survey of the U.S. Bureau of the Census. A description of the basic algorithm and of the extensions which we have developed appears in section 2.

Sections 3 and 4 describe the evaluation studies and the stratification eventually adopted in the two main types of area distinguished by the LFS sample design, namely non-self-representing units (NSRUs) and self-representing units (SRUs). Section 4 also describes how the algorithm was adapted to delineate the primary sampling units (PSUs) within the NSR strata.

Section 5 concludes with a number of observations on the possibility of adapting the new system to other applications.

2. STRATIFICATION ALGORITHM

The basic algorithm used for stratification is a non-hierarchical multivariate algorithm developed by Friedman and Rubin (1967). This choice is based on the results of studies

¹ J.D. Drew and Y. Bélanger, Census and Household Survey Methods Division, P. Foy, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

performed by Judkins and Singh (1981) and Kostanich, Judkins, Singh and Schantz (1981), who assessed a number of stratification algorithms for the Current Population Survey of the U.S. Bureau of the Census.

The latter modified the objective function of the algorithm for sampling with probability proportional to size (PPS), and we have added the capacity to formulate compact, contiguous strata. A more complete description of the following appears in Foy (1984).

2.1 The objective function of the algorithm

The algorithm is designed to partition the stratification units (census enumeration areas) into strata which are as homogeneous as possible with respect to a number of variables of interest that is, by minimizing the sums of the squares within each stratum.

The expressions for the sums of squares in the case of sampling with PPS are shown below after introduction of the following notation:

- L = number of strata to form
- N = total number of units (enumeration areas)
- N_k = number of units in group (stratum) k ; ($N_1 + N_2 + \dots + N_L = N$),
- T_{jk} = size measure of unit j in group k ,
- $T_{.k}$ = size measure of group k ,
- $T_{..}$ = total size,
- ${}_iX_{jk}$ = observed value of variable i for unit j in group k ,
- ${}_iX_{.k}$ = total observed values of variable i in group k ,
- ${}_iX_{..}$ = total observed values of variable i ,
- W_i = weighting factor of variable i (see section 2.4 for further details),
- p = number of variables of interest.

Thus, the expression of the total sum of squares with PPS, of variable i is given by

$$SCT_i = \sum_{k=1}^L \sum_{j=1}^{N_k} \frac{T_{jk}}{T_{..}} \left(\frac{T_{..}}{T_{jk}} {}_iX_{jk} - {}_iX_{.k} \right)^2.$$

This is also the variance expression of the estimate of ${}_iX_{..}$ when a unit is selected with PPS. The total sum of squares weighted for all variables is thus

$$SCT = \sum_{i=1}^p W_i SCT_i.$$

The within-group and between-group sums of squares are obtained respectively by the following expressions:

$$SCW_i = \sum_{k=1}^L \frac{T_{..}}{T_{.k}} \sum_{j=1}^{N_k} \frac{T_{jk}}{T_{.k}} \left(\frac{T_{.k}}{T_{jk}} {}_iX_{jk} - {}_iX_{.k} \right)^2$$

and

$$SCB_i = \sum_{k=1}^L \frac{T_{.k}}{T_{..}} \left(\frac{T_{..}}{T_{.k}} {}_iX_{.k} - {}_iX_{..} \right)^2.$$

Their sums of squares weighted for all variables are given respectively by

$$SCW = \sum_{i=1}^p W_i SCW_i$$

and

$$SCB = \sum_{i=1}^p W_i SCB_i.$$

The within-group sum of squares of variable i , SCW_i , is also the variance expression of the estimate of ${}_iX_{..}$ when a stratum, and subsequently a unit of this stratum, is selected with PPS.

Once again, we have the following result:

$$SCT_i = SCW_i + SCB_i, \quad (i = 1, \dots, p)$$

and

$$SCT = SCW + SCB.$$

The objective function of the stratification program is SCW , the within-group sum of squares weighted for all variables. We define the stratification index for variable i , I_i , as:

$$I_i = 100 \times \frac{SCB_i}{SCT_i} \quad i = 1, \dots, p.$$

A high index value indicates a good clustering.

2.2 Identification of the Best Clustering

One way of identifying the best clustering would be to generate all the possible partitions of N units into L groups and then simply select the one which minimizes the objective function. This approach is rarely feasible because the number of possible partitions may be unmanageably large.

Friedman and Rubin (1967) suggest the following algorithm. Begin with any partition of the N units into L groups. Consider moving a single unit to a group other than the one it is in. Move the unit to the group which offers the greatest reduction in the objective function. If no move will produce a reduction, leave the unit where it is. Using the partition thus created, we process the second unit in the same way, then the third, etc. The application of this procedure to each unit becomes an iteration which the authors describe as a *hill-climbing pass*. After several hill-climbing passes, the algorithm reaches a point at which no move of a single unit will produce a reduction in the objective function. This point is described as a local minimum of the objective function because it is dependent on the starting partition. Another starting partition might have achieved an even lower value of the objective func-

tion. To move beyond the local minimum, Friedman and Rubin describe two procedures, the *forcing pass* and the *reassignment pass*. By applying their algorithm to data described in their article, they obtain the highest known value of the objective function 10 times out of 14 runs from different starting partitions. They use another objective function, which is maximized. With some less well-structured data, the highest value was reached in 3 out of 11 runs, although it is impossible to be certain that this is the optimal solution. In their opinion, the forcing pass and reassignment pass methods are useful only on occasion. They have more confidence in the results obtained through the use of a number of starting partitions. This view is supported by Judkins and Singh (1981). We therefore decided to use the technique involving a number of starting partitions.

Because the algorithm moves only one unit at a time, calculation of the objective function is simplified. Following the initial calculation of the objective function, we merely recalculate the contribution to the objective function of the two groups involved in the move of the unit in question.

2.3 Contiguity

Previous LFS sample designs have used strata composed of contiguous geographic units; that is, each unit in a given stratum had to be touching at least one other unit in the same stratum. One of the main reasons was the assumption that such strata would retain the efficiency of the sample design for a longer period of time than if they were formed of discontinuous units.

In order to assess this assumption and to adopt the best possible stratification, we considered two means of taking geography into account in the stratification. The first method is described by Dahmström and Hagnell (1978), and consists of the use of centroids as variables of interest. This method uses two geographic variables (centroids), which are transformations of longitude and latitude. It yields compact strata, that is, strata in which the distance between units is made minimal by minimizing the usual within-group sum of squares of the centroids. However, the minimization is tempered by minimization of the other variables of interest. Moreover, there is no assurance that these strata will be composed of contiguous units.

The other method, which we describe as the contiguity vectors approach, is new. It guarantees contiguous, but not necessarily compact, strata. Studies described in section 3 dealt with the use of each of these methods in isolation or in combination.

2.3.1 Contiguity Vectors

To ensure the formation of contiguous strata, we proceeded as follows. Optimization is performed as described in the preceding section but beginning, in this case, with a starting partition which is contiguous, and permitting the movement of unit j from stratum A to stratum B only if, in addition to reducing the sums of squares, the following conditions are met:

- i) unit j is contiguous to a unit in stratum B
- ii) the movement of unit j to stratum B will not disrupt the contiguity of stratum A .

In order to verify these two conditions, it is essential that we know the links of contiguity between the units. Consequently, each unit must be assigned a contiguity vector containing a list of the units contiguous to it.

The first condition is easy to verify. In order to ensure that unit j is contiguous to a unit in stratum B , we must simply find one unit in its contiguity vector which is in stratum B .

The second condition is more difficult to verify. The principle is that a stratum is said to be contiguous if each pair of units in that stratum can be connected by a contiguous chain of units in that stratum. Suppose we want to move unit j from stratum A to stratum B . We therefore have to find, for each pair of units in the contiguity vector of unit j within stratum A , another link from among the units of stratum A . At this stage, the problem becomes like finding a path through a maze.

An algorithm has also been designed to create random starting partitions whose strata are contiguous.

2.4 Weighting of Variables

The weighting factors are of particular importance, since they determine the contribution of each variable to the cluster analysis.

It is usually preferable to standardize the variables by making the weighting factors inversely proportional to the total sum of squares of each variable. This standardization makes it possible to obtain a comparable contribution by each variable to the cluster analysis.

If, after standardization, we want to assign one or more variables greater importance in relation to the other variables in the optimization, we can do so by specifying a weight greater than 1 (normal). For example, a variable with a weight of 2 would have double importance. As described in section 3.2, we tested a number of combinations of weights for the geographic and non-geographic variables in an effort to obtain compact strata without unduly affecting the minimization of the other variables.

3. STRATIFICATION IN NON-SELF-REPRESENTING UNITS

3.1 Old Design (Platek and Singh 1976)

For the purposes of the LFS, each of Canada's ten provinces is divided into a number of economic regions (ERs), consisting of areas having similar economic structures. The boundaries of the ERs are determined in consultation with the provinces. These ERs are used as primary strata. The next stage in stratification is the partition of each ER into self-representing units (SRUs) and non-self-representing units (NSRUs). The self-representing units are cities in which the expected sample is large enough to represent at least one interviewer assignment; the NSR part make up the rest of the ER. Different sample designs are used in the SRUs and the NSRUs, because the population in the NSRUs is much more widely dispersed, necessitating a larger number of sampling stages. For the same reasons, we are retaining the concept of the SRUs and the NSRUs in the redesign.

In the old design, the NSR portion of each ER was stratified into a maximum of 5 contiguous strata with a population of between 36,000 and 75,000, based on the main characteristics of the 1971 census population, as described below and as discussed at greater length by Platek and Singh (1976).

The labour force was divided into 7 categories by industry. In each ER, the three largest industries were selected on the basis of specific criteria. The unit chosen for stratification was the combined municipality, which is the geographic region enclosed within a rural municipality and as such, often contains within its boundaries urban municipalities which are geographically smaller. By comparing, for each of these units, the proportions of the labour force working in each of the three categories with the corresponding proportions at the ER level, we identified the units showing a certain similarity which were grouped into strata. This comparison was done visually with graphics. Adjustments were occasionally necessary to satisfy the size and contiguity constraints.

Within each stratum, 12 to 15 PSUs were formed, all of them representative of the stratum in terms of the stratification variables, and of the ratio of rural to urban population. The rural parts of the PSUs were formed of contiguous EAs, and the urban parts were chosen to be as near to the rural part as possible. The sizes of the strata and the PSUs were determined so that, with two PSUs per stratum, the expected sample was equivalent to one interviewer's assignment size. On the basis of these criteria, and depending on the province, the population of the PSUs varied between 3,000 and 5,000 persons. Within the PSUs, sampling occurred in 2 or 3 stages.

3.2 Studies on Stratification during Redesign

Our studies were designed to produce conclusions which would assist in certain decisions relating to the following aspects of stratification: variables to be used, types of strata (wholly rural, wholly urban, or mixed), and the importance to be assigned to contiguity. Given the very limited time available for studies prior to the formation of the new strata and PSUs, and the general expectation that contiguous strata would be preferable over time to discontinuous strata, the first two aspects were given priority.

Some experimenting was required to find the best means of achieving contiguity, either by contiguity vectors, centroids or a combination of the two. However, following the redesign, a more detailed study was undertaken on the relative desirability of contiguous versus discontinuous strata.

3.2.1 Study on Variables and Type of Stratification

One constraint on the stratification method used in the old sample design was the limited number of stratification variables which could be taken into consideration (3 per ER).

With the new algorithm, this constraint is eliminated. In addition to the seven industry variables, we wished to determine the effect caused by the use of variables relating to the survey topic, such as employment, unemployment and income, and by such characteristics as education, housing and population. The latter characteristics have proven extremely efficient in similar studies performed by the U.S. Bureau of the Census for the Current Population Survey.

Table 1 describes the various options studied with respect to the choice of variables.

As regards the type of stratification, it was decided to study the effect of having separate strata for the rural and urban parts of the ERs, as an alternative to the mixed method of the old design.

The constraints on the sample design requiring PSUs to be approximately equivalent in population size, and the ratio between rural and urban population to remain generally the same for each PSU, frequently resulted in a lack of contiguity between the rural and urban parts of the PSUs. This led to an erosion in the presumed correspondence between the PSU and the interviewer assignment. Stratification into separate rural and urban parts, which could be substratified on an optimal basis, was, it was felt, a possible solution to this problem.

The study dealt with 11 economic regions from across Canada. The strata were defined on the basis of 1971 Census data, and assessed on the basis of 1981 census data. In performing the stratification, we used the 1971 Census enumeration areas as our stratification unit, except in Quebec and Ontario. For these two provinces, we selected census subdivisions, since the large number of EAs in certain ERs (up to 400) would have made execution of the computer programs extremely costly.

We used a conversion file between the geographic units of the two censuses to perform the evaluation based on the 1981 Census. The indices based on the 1981 data were considered more appropriate for evaluation purposes, since in fact the stratification data will be an average of 7 or 8 years old for the life of the sample design. Table 2 shows the indices based on both 1971 and 1981 census data.

Table 1
Stratification Options by Variables

Variables	Stratification option				
	1	2	3	4	5
Industries (7) ^a	x	x	x	x	x
Income		x	x	x	x
Employed		x	x		x
Unemployed		x	x ^b		x
Demography (2) ^c				x	x
Housing (4) ^d				x	x
Education (1) ^e				x	x

^a number of persons employed in agriculture, forestry and fisheries, mines manufacturing, construction, transportation, services.

^b double weighting on unemployment.

^c population 15-24, population 55 and over.

^d 1-person households, 2-person households, owned dwellings, total gross rent.

^e secondary education.

For this study, we chose to form contiguous, compact strata, using contiguity vectors and centroids with an average weight of three (see subsection 3.2.2). The number of strata per ER was the same for all options.

The following conclusions were drawn from the results of the study, which are summarized in table 2.

Type of Stratification: Rural/urban stratification was far superior to total stratification in the case of the *agriculture* variable, which is not surprising. The same phenomenon was evident for the *manufacturing* variable, although it was less spectacular. For the *income* variable, rural/urban stratification was also initially more satisfactory, but it was not particularly robust (that is, the index deteriorated over time). Rural/urban stratification was preferable for the *unemployed* variable, while there was little difference for *employed*.

Stratification Variables: Option 4, in combination with rural/urban stratification, was clearly superior for the *unemployed* variable. As regards the other variables, option 5 was slightly more satisfactory than the rest for *employed* and *income*.

3.2.2 Study on Contiguity

As previously mentioned, it was decided to retain the concept of contiguous strata for the LFS. Such strata should be better for the production of small area estimates, because of their better geographic representation. In addition, it was felt that contiguous strata would maintain the efficiency of the sample design for a long period of time.

Table 2
Stratification indices for Option

Stratification variables	Total		Rural/Urban	
	1971	1981	1971	1981
Unemployed				
7 industries	5.4	0.1	9.9	3.8
7 industries + income + employed + unemployed	5.2	2.3	10.2	3.4
7 industries + income + employed + unemployed $\times 2$	7.4	2.3	10.2	5.3
17 variables	6.3	6.4	11.3	4.7
15 variables (excluding employed + unemployed)	3.6	0.1	9.8	9.0
Employed				
7 industries	2.9	0.5	8.9	4.8
7 industries + income + employed + unemployed	8.8	2.7	8.6	3.2
7 industries + income + employed + unemployed $\times 2$	9.1	2.8	13.1	2.2
17 variables	14.1	7.8	12.2	6.4
15 variables (excluding employed + unemployed)	6.3	1.6	11.4	3.7
Income				
7 industries	7.4	5.7	18.9	9.5
7 industries + income + employed + unemployed	11.2	6.8	22.1	5.9
7 industries + income + employed + unemployed $\times 2$	10.3	6.8	28.3	9.5
17 variables	10.5	9.4	24.4	11.9
15 variables (excluding employed + unemployed)	21.0	5.3	28.9	4.5
Agriculture				
7 industries	7.4	9.7	37.0	26.0
7 industries + income + employed + unemployed	7.6	7.8	40.0	28.7
7 industries + income + employed + unemployed $\times 2$	8.6	7.9	43.2	31.0
17 variables	6.1	1.1	40.3	31.8
15 variables (excluding employed + unemployed)	7.0	0.4	42.7	29.0
Manufacturing				
7 industries	14.7	8.5	16.9	13.2
7 industries + income + employed + unemployed	10.9	6.6	16.5	12.1
7 industries + income + employed + unemployed $\times 2$	5.5	4.3	14.8	16.1
17 variables	12.5	13.5	13.3	10.7
15 variables (excluding employed + unemployed)	7.2	1.4	14.1	16.4

The next question was how to use the centroids or contiguity vectors, or a combination of the two, to obtain compact, contiguous strata without allowing the geographic constraints to affect minimization of the other variables unduly.

The study was performed with the same 11 economic regions. As anticipated, the use of contiguity vectors alone resulted in strata which were contiguous, but often irregular in shape. At the same time, the use of centroids alone, even with high weights, failed to provide any guarantee of absolute contiguity.

By varying the weight of the centroids relative to the other variables, we found that a combination of a centroid weight of 3 and contiguity vectors offered a good compromise between compactness and non-geographic optimization.

3.3 Design Stratification

In view of these results and the superior results shown by a sample design using rural/urban stratification in a study on cost variance optimization (Choudhry, Lee, Drew 1985), we decided to use separate stratification for all economic regions except those in which either the rural or the urban population was too small to form at least one stratum. It was determined that each stratum should provide a sample of at least 90 dwellings, corresponding to the selection of two PSUs with a minimum take of 45 dwellings each. In cases where this requirement could not be met, we decided to proceed with overall stratification and thus to form mixed strata. This criterion led to the adoption of separate strata in over 2/3 of the ERs.

As regards the stratification variables, we compromised on a stratification based on the 15 variables of option 4 plus *employed*. *Employed* was added because its inclusion in option 4, as compared to option 5, improved the performance of the *employed* and *income* characteristics. For the same reason, *unemployed* was excluded as a stratification variable.

For the geographic constraints, it was decided to use the contiguity vectors in combination with a uniform centroid weighting of 3 in all economic regions.

A decision was also required as to the number of strata per ER. In practice, in most of the cases, there was no choice. According to the sample design, each PSU corresponds to one interviewer assignment, and we wanted to select at least two PSUs per stratum in order to produce unbiased variance estimates. Given these constraints, in almost 2/3 of the cases, only one stratum was formed with 2 or 3 selections, in the urban or rural parts or a combination of the two. In the other cases, stratification was performed in such a way as to permit the selection, again, of 2 or 3 PSUs per stratum. This decision was based on another study showing slight reductions in variance with this approach, as compared to the old sample design in which 4 to 6 PSUs were selected from each stratum (Choudhry, Lee, Drew 1985).

3.4 Study on Robustness of Contiguous and Discontiguous Strata

Robust strata are strata that maintain the efficiency of the sample design over time. Following redesign, a study was performed to determine whether contiguous strata would be more robust over time, as had been hypothesized.

The study dealt with three economic regions in Ontario, ERs 520, 540 and 580 (1981 numbering). For each of these regions, the results of the new stratification (selected for the redesign of the LFS), which consists of contiguous strata, were compared with a stratification without contiguity constraints. The strata were defined on the basis of the 1981 data, and evaluated on the basis of the 1971 data. For the contiguous strata, we used contiguity vectors with centroids, while for the discontiguous strata, we tested two options using centroid weights of 0 and 3 respectively. The stratification variables used were the same 16 variables described above (modified option 4).

The results are shown in table 3. We see that in general, the total index calculated on stratification is higher for the two options in which contiguity is not necessary, as might be expected (1981 column). However, these two options also give higher indices over time (1971 column).

Do we really need contiguous strata? Before answering this question, we would have to perform a more in depth study involving ERs from a number of provinces. Evaluation of stratification robustness would however pose certain problems. It is easy to evaluate robustness in Ontario, since stratification there is performed at the census subdivision level, which has changed very little since 1971. When stratification is performed at the level of the enumeration areas, which are very changeable, it is extremely difficult to obtain precise figures on robustness when the strata are neither compact nor contiguous.

However, should it prove that stratification without contiguity is more satisfactory, this could compensate for the possible problems involved in production of small area estimates. It could also open new horizons: once contiguity constraints are eliminated, why could we not begin by forming compact, but not necessarily contiguous, PSUs, and then grouping them into strata? This question also could only be answered by further and more detailed studies.

3.5 Formation of PSUs

The clustering algorithm was modified to permit PSU delineation in rural and mixed strata. In the rural strata in particular, the formation of the PSUs is conceptually very similar to stratification. The only difference relates to the fact that in stratification, we attempt to minimize the sums of squares of the geographic and non-geographic variables within each stratum, while in PSU formation, we want to minimize the sums of squares of the geographic variables (to obtain compact PSUs in order to reduce costs) and to maximize those of the non-geographic variables. The latter criterion enables us to obtain PSUs which are as heterogeneous as possible in terms of characteristics, so that they are all properly representative of the stratum during sampling.

There is, however, a conflict between the desired compactness of the PSUs and their heterogeneity, because of the tendency of adjacent units to possess similar characteristics. Because of low computer costs, we performed 3 delineations per stratum with centroid weights of 10, 15 and 20, relative to the other variables. The results of each delineation were then plotted on a graph whose axes are the centroids (see Figure 1). We then selected the best of the 3 delineations on the basis of the quality of variable optimization, as reflected by the stratification indices, and through reference to the graphs. A compactness index was also taken into consideration. In most cases, as it worked out, a centroid weight of 10 or 15 was selected.

Table 3
Stratification Indices by Geographic Constraints

Economic Region	No. of Strata	Geographic Constraints					
		Contiguity and Centroids (weight of 3)		Centroids (weight of 3)		None	
		1981	1971	1981	1971	1981	1971
520	2	32.2	28.5	30.2	30.1	34.5	27.0
540	3	21.8	14.1	24.9	17.8	35.2	26.8
580	4	22.8	18.9	41.4	33.7	43.7	38.5

Formation of the PSUs in the mixed strata led to an additional constraint. We wanted the proportion of urban population to be approximately the same in each PSU. Since we also wanted the PSUs to have approximately equal total populations, it was therefore necessary in some cases to split the large urban centres among a number of PSUs. The following solution was adopted:

1. The average number of parts of urban centres which a PSU will receive (N) is determined. This number depends on the proportion of the urban population in the stratum and on the number of urban units. In practice, it was set at 1 or 2. Certain strata without sufficient population or a sufficient number of urban units were reclassified as entirely rural strata.
2. The number of parts into which each urban centre will be divided is determined. The total number of parts must equal N times the number of PSUs and each urban centre is divided into a number of parts proportional to its population.

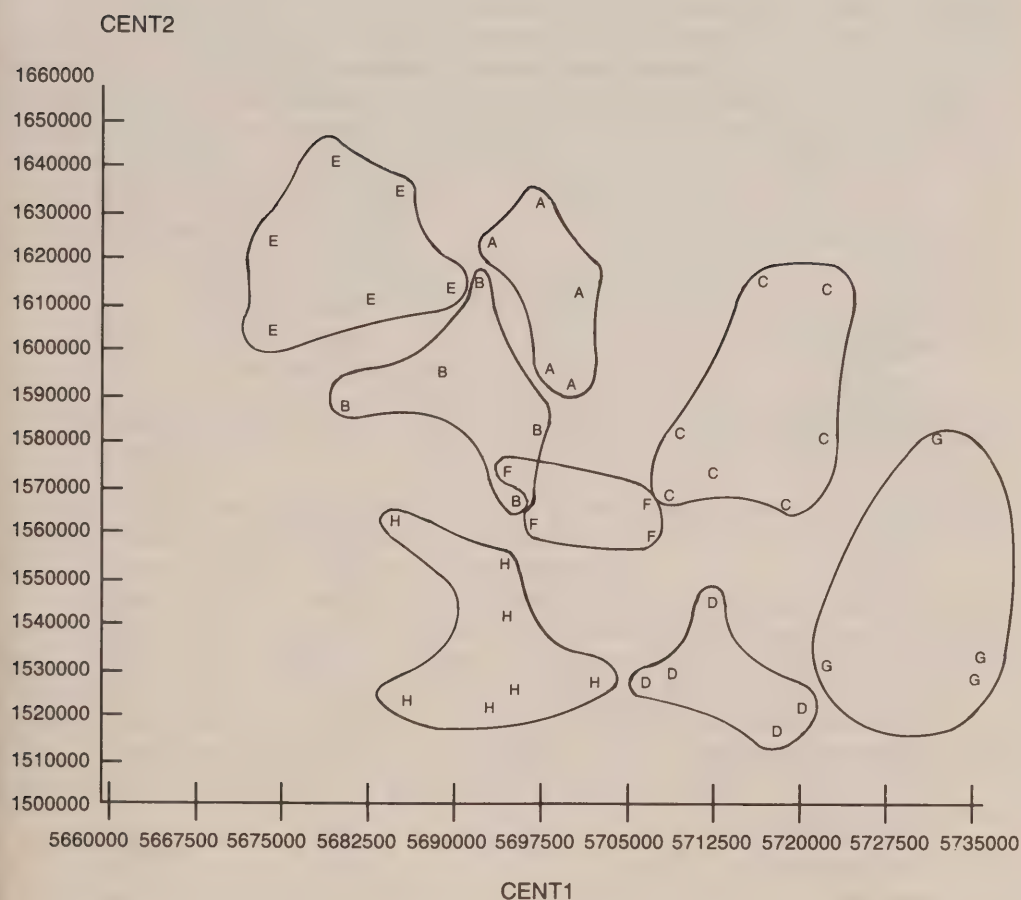


Figure 1. Example of PSU Delimitation. Each stratification unit is represented by a letter identifying the PSU to which it belongs. The PSUs are circled for clearer differentiation.

3. The optimal stratification program is applied, considering each part of an urban centre as a distinct stratification unit and adding the *urban population* variable to the other stratification variables. The weight assigned to this variable is adjusted to obtain the most evenly balanced rural/urban distribution possible within each PSU, without unduly disrupting compactness and overall optimization. This can be done only by trial and error. In practice, we found that a weight of 10 or 15 on urban population, relative to the other variables, produced satisfactory results.

In the urban strata, the PSUs were composed of urban centres. In some cases, small centres, relatively close together, were combined, without considering characteristic optimality.

Table 4 gives the average delineation indices for the PSUs in rural, mixed and urban strata. For the non-geographic variables, the lowest index represents the best delineation, while the opposite is true for the centroids. The results are clearly better, in terms of characteristic optimality, for the rural and mixed strata, in which the clustering algorithm was used. The high indices of the centroids show that the PSUs are relatively compact.

Table 4
Average PSU Delineation Indices

Variables	Type of stratum		
	Rural	Mixed	Urban
Agriculture	8.1	8.3	9.0
Forestry	21.8	24.5	35.9
Mines	20.6	36.0	57.0
Manufacturing	15.1	22.9	53.3
Construction	9.0	11.4	22.7
Transportation	9.9	12.8	22.7
Services	9.4	12.8	29.1
Employed	7.7	10.2	23.6
Unemployed ^a	13.6	14.2	18.6
Income	8.9	11.2	23.7
Population 15-24	9.4	13.4	29.8
Population 55+	7.4	13.9	34.5
1-person households	5.1	7.4	13.0
2-person households	7.9	11.9	28.1
Owned dwellings	6.8	12.5	29.4
Total gross rent	5.1	7.7	14.4
Secondary education	9.1	10.5	17.4
Total population ^a	3.2	4.0	10.5
Dwellings ^a	5.9	8.9	18.6
Centroid 1	91.6	92.7	99.2
Centroid 2	90.5	91.7	97.2

^a Not used as a variable in optimization.

4. STRATIFICATION IN SELF-REPRESENTING UNITS

4.1 Old Design

The self-representing units of the old sample design corresponded to those cities large enough to yield an expected take equivalent to one interviewer assignment. The lower limit for SRUs varied from 10,000 persons in the Atlantic provinces to 29,000 in Quebec and Ontario.

The large SRUs were geographically stratified by grouping 3 to 5 contiguous census tracts (CTs), without any attempt to optimality. CTs are geostatistical units with populations between 3,000 and 5,000; because of their stability from one census to the next, they are practical operational units. It was felt that these strata would be efficient in estimating characteristics, and that their small size (between 10,000 and 15,000 persons) would permit sample updating in areas experiencing rapid growth, without disrupting the rest of the sample.

In addition to the area frame, an open-ended frame was set-up for apartment buildings in the large cities.

4.2 Study on stratification

Three large SRUs were considered in this study, namely Quebec City, Ottawa and Toronto. The stratification unit selected was the census tract. Because of operational constraints imposed by the stratification program, it was necessary to break Toronto up into six parts, corresponding generally to the city's major natural divisions. Stratification was carried out separately in each of these parts. The same 16 stratification variables finally selected in the NSR part were used.

Two main options were evaluated:

Option 1: Two-level stratification:

- contiguous, compact primary strata, with a centroid weighting of 3 and an expected take of approximately 150 dwellings.
- secondary strata - 4 or 5 per primary stratum, formulated without geographical constraints.

Option 2: compact stratification formulated with the use of centroids (weight of 3) and without contiguity vectors, comparable in size to the secondary strata of option 1.

Table 5 shows the results of the comparison between the old stratification and the two options studied. As in the NSR part, the strata were defined on the basis of 1971 Census data, and then evaluated on the basis of 1981 data.

We see that the two options studied consistently show better indices than the old stratification, with the possible exception of the first three variables, which, in any case, are of limited importance in cities. The old stratification nevertheless performed quite well, considering that it was carried out without any concern for optimality.

We also note that all three methods provide generally robust stratification over time, as reflected by the comparison between the indices for 1981 and 1971. Major exceptions to this rule, unfortunately, appear to be the employed and unemployed characteristics.

4.3 New Design

Given the similarity in results between the two options studied, it was decided to adopt two-level stratification (option 1) in large cities where the sample consists of 300 or more households, for the following reasons:

- i) Contiguity in the primary strata gives us a suitable unit for sample updating.

- ii) The primary strata can be used for the formation of interviewer assignments. The size of the strata was determined so that the sample within the geographic area, that is, the area frame sample plus the sample for the apartment frame, corresponds to two interviewer assignments (160 households in the city core and 120 elsewhere).
- iii) Two-level stratification leads to better representation of the correlated response variance in variance estimates. In the old sample design, there was usually only one interviewer per stratum, resulting in an underestimate of this component of the variance. With non-geographic secondary strata, but geographic interviewer assignments, this problem will be less frequent.

The cost constraints associated with the computer time involved forced us to deal with certain SRUs on an individual basis. In fact, the Montreal region was divided into seven independent parts, during stratification. The same was done with Toronto (5 parts), Winnipeg (2 parts), Calgary (2 parts), Edmonton (2 parts) and Vancouver (3 parts). These divisions were made on the basis of natural criteria as suggested by the geography of these regions.

In large SRUs, apartment buildings existing at the time the sample design was developed were sorted by the primary strata in which they were physically located in order to achieve an implicit stratification of this sample.

Table 5
Comparison of Three Stratification Methods (SRUs)

Variables	Old Design		Two-level Stratification (Option 1)		Compact Stratification (Option 2)	
	1971	1981	1971	1981	1971	1981
Agriculture	5.5	2.9	3.2	1.8	3.4	1.8
Forestry	2.2	2.3	2.1	1.7	2.2	2.3
Mines	7.6	4.9	8.5	4.1	7.6	4.0
Manufacturing	34.7	35.0	36.6	34.1	39.1	35.0
Construction	32.5	29.6	39.7	30.1	42.4	33.4
Transportation	9.2	6.8	18.0	11.6	20.0	11.6
Services	29.5	27.5	45.8	33.1	46.7	32.1
Employed	15.1	8.0	31.4	14.1	32.8	12.6
Unemployed ^a	14.6	5.7	14.9	6.7	15.5	7.1
Income	39.4	38.6	51.8	29.8	53.6	48.0
Population 15-24	9.6	15.2	12.5	17.5	13.3	14.9
Population 55 +	27.9	18.3	34.0	20.8	32.6	18.5
1-person households	20.3	19.2	36.3	33.8	37.8	35.0
2-person households	21.9	20.3	40.3	30.9	40.1	30.2
Owned dwellings	20.3	15.3	29.7	22.9	32.1	24.9
Secondary education	32.6	42.4	50.3	47.9	51.6	49.1
Population 15 + ^a	27.0	8.2	38.0	13.4	37.6	12.0
Dwellings ^a	21.8	18.5	41.7	33.8	42.1	34.3

^aNot used as a stratification variable.

In medium-sized SRUs, where the sample was not large enough to justify two-level stratification, optimal strata were simply constructed by means of the stratification program, without the application of geographic constraints.

The smallest SRUs, those not broken into block faces for census purposes, were manually stratified, without any attempt at optimality.

Finally, we might note that the phase-in period of the new sample produced a further constraint. For large SRUs, core areas were defined as consisting of complete old-design strata that were unaffected by boundary changes. By having strata in the new design respect these core areas, we ensured that during phase-in, the new sample in core areas represented the same geographic area as the old, which permitted gradual replacement of the old sample by the new without the need for a costly parallel build up of new sample (Mayda, Drew, Lindeyer 1985).

5. CONCLUSIONS

Use of multivariate clustering algorithm enabled us to develop a very general stratification, thus strengthening the LFS in its role as a general household survey. In addition, automation of the various stages of stratification in the NSR and SR parts, and delineation of the PSUs in the NSRUs, led to a significant reduction in the cost and time required to redesign the sample.

The system is documented (Foy 1984) and can be used for the stratification of other surveys. It may also be used in situations requiring the definition of statistical or administrative regions, using a full range of variables.

For the LFS, one aspect requiring further research relates to the selection of contiguous or discontinuous strata, and the implications of discontinuous strata on sample design.

ACKNOWLEDGEMENTS

The authors would like to thank Sylvie Trudel and Marc Joncas for their assistance in carrying out the studies mentioned in this report, and the members of the LFS Sample Redesign Committee for their valuable suggestions. They are also grateful to the referee for his helpful comments.

REFERENCES

- CHOUDHRY, G.H., LEE, H., and DREW, J.D. (1985). Cost-variance optimization for the Canadian Labour Force Survey. *Survey Methodology*, 11, 33-50.
- DAHMSSTRÖM, P., and HAGNELL, M. (1978). The formation of strata using cluster analysis. Internal document, Department of Statistics, University of Lund, Sweden.
- FOY, P. (1984). Stratification program for the Canadian Labour Force Survey: User's guide. Internal document, Census and Household Survey Methods Division, Statistics Canada.
- FRIEDMAN, H.P., and RUBIN, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.
- JUDKINS, D.R., and SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *American Statistical Association Proceeding of the Section on Survey Research Methods*, 274-284.

- KOSTANICH, D., JUDKINS, D.R., SINGH, P.R., and SCHANTZ, M. (1981). Modification of Friedman-Rubin's clustering algorithm, for use in stratified PPS sampling. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 285-290.
- MAYDA, F., DREW, J.D., and LINDEYER, J. (1985). Phase-in of the redesigned Labour Force Survey. Internal document, Census and Household Survey Methods Division, Statistics Canada.
- PLATEK, R., and SINGH, M.P. (1976). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.
- SINGH, M.P., DREW, J.D., and CHOUDHRY, G.H. (1984). Post '81 censal redesign of the Canadian Labour Force Survey. *Survey Methodology*, 10, 127-140.

Sampling Microfilmed Manuscript Census Returns

D.R. BELLHOUSE¹

ABSTRACT

In the first part of the paper a review of the historical literature concerning microfilmed manuscript census records is given. Several types of sampling designs have been used ranging in complexity from cluster and stratified random sampling to stratified two-stage cluster sampling. In the second part, a method is given to create a public use sample tape of the 1881 Census of Canada. This work was part of a pilot project for Public Archives of Canada and was carried out by the Social Science Computing Laboratory of the University of Western Ontario. The pilot project was designed to determine the merit and technical and economic feasibility of developing machine readable products from microfilm copies of the 1881 Census of Canada.

KEY WORDS: Computerized random sampling; Microfilmed records; Multi-stage designs; Public use samples; Stratification.

1. INTRODUCTION

To write a history of any person or people the historian must rely on the applicable source material. Many historians today seek to write a history of the *common man*. In this area of historical research the source material may include items such as census returns, land records, and business directories. This paper focuses on the use of census returns as a source material. The major problem with using census data is that there is a large mass of it. For an historian with a reasonable research budget there is not enough money, time or manpower to sift through all the census returns. The solution is to take a random sample of the returns. Most census returns available to the historian are microfilm copies of the returns. In Canada this includes the colonial censuses of 1841, 1851, and 1861 and the Census of Canada for 1871 and 1881. The problem then becomes one of finding the appropriate design to sample returns from the microfilm copies.

In section 2 of the paper a review of sampling techniques that have been used by historians is given. The use of sampling techniques by historians has been very uneven. Some applications have been very good; the use of a particular technique was well thought out and applied. At the other end of the spectrum other historians appear to have used overly complex designs when it was not necessary. A complex design could lead to design effects much different from 1 which, in turn, could lead to problems in the analysis of the data. See, for example, Rao and Scott (1981) and Holt *et al.* (1980) for discussions concerning categorical data analysis and Scott and Holt (1982) for regression analysis. One other problem with many of the surveys reviewed here is that there is insufficient discussion in the survey report to ascertain the reasons why a particular design was chosen.

In section 3 of the paper a method is given to sample the returns of the 1881 Census of Canada for the purpose of creating a public use sample tape. The work was carried out as part of a project for Public Archives of Canada. The contract for the research was awarded to the Social Science Computing Laboratory of the University of Western Ontario. A description of the sampling design is given here; a complete report of the project is found in Mitchell *et al.* (1982). In some ways the design is similar to the ones used for creating public

¹ D.R. Bellhouse, Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada N6A 5B9.

use sample tapes for the 1971 and 1976 Censuses of Canada. The sampling designs are all based on stratification; however, in the case of the 1881 Census, stratification could only be carried out on a geographical basis.

2. HISTORICAL REVIEW

The sampling literature for historical census documents may be categorized by the type of sampling method that was used. The order of categorization followed here will be in approximately increasing complexity of the sampling design.

2.1 Cluster Sampling

Ornstein and Darroch (1978) have given a simple cost efficient method of sampling and linking census records over time. The heart of the scheme is to form clusters of surnames and then to sample clusters. The clusters are defined by the first letter of the surname. If the same clusters are sampled over various censuses then an individual who appears in more than one census will be in the chosen sample. This reduces the number of cases to be examined for linkage purposes and hence reduces the cost. This design is particularly useful for historical studies of migration or historical changes over time.

2.2 Stratified Sampling

In all of the designs considered here that used stratification, no attempt was made to use optimal allocation. This was because prior knowledge of the variation within strata was not available to any of the researchers. To obtain the required information would have increased the cost of each project substantially.

Hammarberg (1971) used a type of two-phase or double sampling technique in an attempt to decrease the bias incurred by sampling from an incomplete set of records. The records, sampled at the second phase, were business directories for nine counties in Indiana. In the first phase of sampling, he sampled from an assumed complete record set, the 1870 United States Census. The sampling method was stratified random sampling with proportional allocation so that the sample is self-weighting. The strata were the nine counties. Two aspects of this study recur in subsequent historical sampling studies. The strata are geographical areas and the sample is self-weighting.

Hammarberg (1971) also used the classical chi-square test of fit on certain variables to see how well his sample data fit known population distributions from the census reports. In many other studies no attempt was made to check the *representativeness* of the sample.

Soltow (1975) used samples from the 1850, 1860 and 1870 United States Censuses to study wealth in the United States. For each census year he selected a sample from each microfilm reel so that the sample is stratified by reels, an approximate geographical stratification. Soltow's design appears to be a type of systematic sampling. To choose a sample he designated a spot on the screen of the microfilm reader and fed the film through the reader. The feeder arm was given successive half-turns until the manuscript census entry at the designated spot on the screen was acceptable. One criterion for sample unit selection was that the entry had to be male aged twenty years or older. Also, persons "with wealth of \$100,000 or more were sampled 40 times more heavily in 1860 than those under \$100,000" (p.5) so that the design is not self-weighting. Although it is not stated, the *oversampling* of wealthy people appears to have been done in order to obtain a reasonable number of them for comparison to the less affluent sections of society. Soltow (1975) also compared his sample results to the published distributions but made no statistical tests for goodness-of-fit. He found that the sample data conformed well to the census results in terms of averages and proportions on

various variables. This was true even for variables such as mean wealth, a result which is surprising in view of the oversampling of the more wealthy individuals and since his estimate appears to be the sample average.

In studying the relationship between ethnicity and occupation, Darroch and Ornstein (1980) used a sample of the 1871 Census of Canada. A description of the sampling method is given in Ornstein (1978). For the purposes of both studies it was necessary to *oversample* some ethnic groups so that the design used was not self-weighting. On ignoring the oversampling of certain ethnic groups, the sampling method used was stratified random sampling. The stratification is based on the geographical hierarchical structure of the census records : provinces, districts within provinces, sub-districts, and divisions within sub-districts. The division corresponds to the modern enumeration area. The natural stratification variable seems to be divisions. However, Ornstein (1978) further subdivided divisions into smaller groups which comprise the strata and then sampled two households per stratum. How the further subdivision was made is not given, but Ornstein states that the reason for further stratification is that sampling two units per stratum minimizes the variance of estimates of certain population values. Although it is not stated, it appears that Ornstein (1978) was trying to increase the efficiency of stratification by forming strata within a division as homogeneous as possible. By stratifying in this way the cost to sample was increased. One other aspect of Ornstein's (1978) method is that it was necessary to make at least two passes through the microfilms, the first to obtain the number of households per division and the second to sample the household.

Johnson (1978b) and Graham (1980) obtained a public use sample of the United States Census of 1900. Johnson (1978a) has described some related work in sampling the 1860 Rhode Island Census schedules. The sample was chosen by obtaining random lines on the microfilm, and then by searching for the chosen lines using a microfilm reader with an odometer attachment. Because of the sample selection procedure, the overall sample size is random. A number of criteria are given in Graham (1980, p. 41) for including or excluding sampled lines. The sampling scheme is stratified random sampling with microfilm reels as strata. The stratification is geographically based provided that the contiguous census returns are all grouped in the same microfilm. The advantages of this scheme are that it is operationally efficient and only one pass through the microfilm is needed. Also, it avoids the problem of empty strata or one unit per stratum when the sampling fraction within a stratum is small. One disadvantage is that, since one pass through the data is made, potentially major problems that arise must be dealt with on an ad hoc basis.

2.3 Stratified Cluster Sampling

Bateman and Foust (1974) obtained a sample of farms in the northern United States from the 1860 United States Census. The north was divided into two strata, East and West, and a random sample of rural counties was chosen in each stratum. Within a county one rural township (the cluster) was chosen at random and information was collected on every farm in the township. One reason for clustering appears to be due to cost considerations. The farms were obtained from the census of agriculture schedules and demographic information on the owners or operators was obtained from the census of population schedules. By remaining in the same township the work of matching farms to owners is minimized. Swierenga (1983) has provided a second reason for cluster sampling. He states that township data made it possible to estimate total factor productivity in agriculture and to identify the entire agricultural workforce, including farm laborers not residing in the 12,000 farms included in the sample (p. 793). Since the clusters, townships, were not chosen by probability proportional to size the design was not self-weighting.

Bateman and Foust (1974) also used some tests to check the representativeness of their sample. As in Hammarberg (1971), they applied the chi-square test of fit to compare sample counts to expected population counts. For continuous variables they used the t-test. The estimates of the mean and variance were the *simple* estimates, not based on the sampling design.

2.4 Stratified Two-Stage Sampling

Hammarberg (1977) used a stratified two-stage sampling scheme to sample households in the 1880 census for Utah Territory. The strata are a fairly complicated amalgamation of five geographical regions in Utah, some counties within populous regions and some large towns. Within each stratum, a sample of towns or wards was chosen. Towns which were already strata were included with certainty. Wards are geographical divisions in the Mormon Church similar to parishes in the medieval Christian church. Then a sample of households was taken from the chosen towns or wards. The sample was self-weighting on the household. The rationale for stratifying on geographical areas, given on page 460 is compelling:

“Because the fundamental organization of the mass of people was conceived geographically, and most institutional records, – both church and secular – were organized to correspond to these areal definitions, a sample of the population on an area-by-area basis is also, in large measure, a sample of the records produced and organized for the population.”

McInnis (1977) also used a stratified two-stage sampling design to obtain a sample from the 1861 Canadian Census. He studied the relationship between the number of children per family and the abundance of land in certain areas. He first stratified approximately 300 townships by their dates of settlement. Then he took a sample of townships within strata and samples of farms within townships. His reason for choosing a two-stage sample appears related to cost. A sampled farm was matched to the entry in the agricultural census. It takes less time and hence costs less to sample a few townships and match records for several farms within a township than to stratify on townships and match this record for a small number of farms. The same argument applies to Hammarberg's (1977) work. He was also linking other records to the sampled household.

2.5 Stratified Two-Stage Cluster Sampling

Smith (1978) used a stratified two-stage cluster sampling scheme to study older Americans in the 1900 United States Census. The strata are described as census regions with the counties within these regions as the primary sampling units. The primary sampling units, counties, are chosen with probability proportional to the size of their population. Within a county, several pages of census returns were sampled. Every individual over the age of 50 on each sampled page was recorded. Cluster sampling was necessary since it was too expensive to identify every individual eligible to be sampled. There is also an attempt to compare some sample distributions to the published census results. The statistic used is the standard test statistic for hypotheses on a single proportion although the data are multinomial.

A second stratified two-stage cluster sample known as the Parker-Gallman sample is described in Foust (1968, ch. 2). This sample was drawn from the 1860 Census of the United States to study the cotton growing regions in the South. The strata were 405 Southern *cotton counties*, those counties which produced 1,000 or more 400-pound bales of cotton in the year preceeding Census day. Within a county a systematic random sample of pages from the manuscript census was chosen; with a selected page a block of five farms was chosen

at random, the block being the cluster. Cluster sampling was used because information on a particular farm had to be accumulated from three different census schedules. The matching of the farms in the schedules was described as *very laborious*. Fogel and Engerman (1974, pp. 22-25) have listed several additional samples related to the Parker-Gallman sample. Bode and Ginter (1984) have criticized the content of the sample.

Of the large number of samples reviewed here, the Parker-Gallman sample and the samples drawn by Bateman and Foust (1974) are the two that have been most extensively studied. Swierenga (1983) has reviewed much of the work based on these samples.

3. PUBLIC USE SAMPLES FROM THE 1881 CENSUS OF CANADA

Early in the 1980's Public Archives of Canada obtained *Schedule 1: Nominal Return of the Living* for the 1881 Census of Canada. The returns were microfilmed and currently copies are available in most academic and many public libraries. After producing the microfilm copies, Public Archives of Canada was then interested in producing a machine readable edition of the entire census and/or a machine readable public use samples similar to the public use samples for the censuses of 1971 and 1976 (see Statistics Canada (1975, 1979) for documentation). The Social Science Computing Laboratory of the University of Western Ontario obtained a contract to perform a feasibility study and the author was asked to design a sampling scheme to construct the public use sample. In this section the proposed design is described. A report of the feasibility study is found in Mitchell *et al.* (1982).

Schedule 1 contains information on each individual on age, sex, country of birth, ethnic origin, occupation, marital status, whether or not the person had certain disabilities. The other seven schedules contain information on industry, agriculture, forestry, fishing, and mining. A brief description is found in *Census of Canada 1880-81* Vol. 1, pp. v-xv.

The basic requirements of the public use samples are briefly described. To conform to the 1971 and 1976 public use samples it would be necessary to have two independent samples, one of households and one of individuals. If production of only one sample is economically feasible, however, the first priority is the household sample. The public use sample of the 1900 Census of the United States, described by Johnson (1978b) and Graham (1980) is a sample of households. Moreover, the household appears to be the most important sampling unit desired by historians. On taking another cue from the sample of the 1900 census, a sample size in the order of one hundred thousand individuals for either the individual or the households sample is desirable. For the 1881 Census of Canada this would result in an approximate 2½% sampling fraction in either sample. Finally a stratified sampling design with proportional allocation with geographical areas as strata for both samples is desirable. This conforms to sampling practice so far in the historical literature and ensures a self-weighting design. Within a stratum the units should be chosen by simple random rather than systematic sampling. Although convenient, Johnson (1978a) has maintained that systematic sampling is not appropriate for manuscript census schedules. Neighbours possess similar characteristics and would never be included together in a systematic sample. Historians may be interested in studying those individuals with like characteristics.

Based on these basic requirements the following sampling scheme was proposed for the household sample. The design suggested was stratified random sampling with census divisions (the modern enumeration area) as strata similar to Ornstein (1978) rather than microfilm reels as used by Johnson (1978b) and Graham (1978). The census divisions provide natural geographical strata. In addition, the households are consecutively numbered on the enumerators lists with twenty-five individuals per census manuscript page. Thus, if one preliminary pass is made through the microfilms the number of households in each stratum could be easily obtained. With a 2 - 2.5% sampling fraction and proportional allocation, sample

sizes of smaller than two households are obtained in divisions (strata) with fewer than approximately one hundred households. In these cases the division should be grouped with geographically contiguous strata. Further stratification beyond the division as in Ornstein (1978) seems unnecessary and would substantially add to the sampling costs.

The sampling process can easily be made part of a computing environment. From the point of view of a coder sitting at a computer terminal with a microfilm reader to one side the sampling process is straightforward. When a coder is sampling a division, he merely presses the appropriate keys identifying the division he wants and the number of the first household to be sampled appears on the terminal screen. The coder then moves the microfilm forward to the appropriate household number. Once the data are entered, a *next* key is pressed and the second household number appears. When the final sampled household from that division is obtained, pressing the *next* key will result in an instruction to pick another division to sample. In some situations there may be missing households. For example, one or more of the enumerators sheets containing 25 names may have been lost. In this case, when a coder, in the process of sampling, encounters a missing household, the household is entered as missing and also any other missing household numbers that the coder may notice. The coder then continues sampling to the end of the division. Since at least one household sampled was missing the coder is instructed to rewind the microfilm and to continue sampling in the division. The main feature for the coder in this set-up is that with the exception of missing data situations the coder need only move the microfilm reel forward.

The computing algorithm behind this sampling method utilizes a file containing information about the divisions or division groupings and Bebbington's (1975) algorithm for drawing a simple random sample without replacement. After the initial pass is made through the microfilms a file is created containing the division identifier and the number of households in the division. If the divisions have been grouped then the size of each is recorded. When a coder identifies a division to be sampled the appropriate file entry is examined and the division size is obtained. The required sample size in the division is the division size times the sampling fraction for the whole survey which yields proportional allocation. Then Bebbington's (1975) algorithm makes a sequential choice of sample units from an ordered list, the list here being the ordered household numbers in a division. Each household number is examined in turn and is selected for or rejected from the sample. When a household number is selected the number is printed to the terminal screen and the selection procedure pauses for data entry. The sample numbers selected will be in increasing order so that a forward search only is necessary on the microfilm.

Sampling collapsed strata or grouped divisions can also be done using this algorithm. Suppose L strata of sizes N_1, \dots, N_L have been grouped into one stratum of size $N = N_1 + N_2 + \dots + N_L$. It is necessary only to use the stratum sizes to obtain the sampled household in each stratum. Suppose in the algorithm units $s(1), \dots, s(n)$ have been chosen for the sample, $1 \leq s(i) \leq N$. If for any i ($i = 1, \dots, n$) $N_1 + \dots + N_{h-1} < s(i) \leq N_1 + \dots + N_{h-1} + N_h$, where $N_1 + \dots + N_{h-1} = 0$ for $h = 1$, the unit $s(i)$ is in stratum h and the household number within that stratum is $s(i) - (N_1 + \dots + N_{h-1})$.

The general sampling algorithm can also be modified to account for missing households. The method described does not require enumerating these missing households prior to sampling. When the sampling of a stratum by Bebbington's (1975) algorithm has been completed, two possibilities arise: no missing households were encountered or some were encountered. In the former situation, there is no problem; the sampling has been completed for that situation. In the second situation, the achieved sample size, say m , is less than the desired size n . To obtain a sample of size n of the existing households, it is necessary to sample $n - m$ additional households. To achieve this, the sampling process for this stratum is started again but a list is created of the sampled and known missing households. Suppose there are M

previously sampled and known missing households ($M \geq n$: a coder may notice and record households that are missing other than those which were chosen for the sample). Define an N -dimensional vector v where the value of the u^{th} entry is u , $v(u) = u$ for $u = 1, \dots, N$. The u^{th} entry is a pointer to the u^{th} household in a division. Now delete all entries in v corresponding to households on the microfilm which are missing or previously sampled and collapse the vector into an $(N - M)$ -dimensional vector w . The values $w(u)$, $u = 1, \dots, N - M$ will contain the household numbers left to sample. In the algorithm, it is necessary only to restate the population size as $N - M$ and the sample size as $n - m$.

A separate and independent sample of individuals can be easily obtained using the household method of sample selection with slight modifications. The key to the modifications is that the pages of the enumerators lists are numbered with 25 names to a page. In the first pass through the microfilms, it is necessary to find the final page number and the number of lines on the last page of each division. On applying Bebbington's algorithm the computer will print the page and line number of the individual sampled.

This method of sample selection has been programmed and tested by the Social Science Computing Laboratory with positive results. For example, in the feasibility study the percentage of time spent searching for sampled units represented approximately 6% of the total estimated data entry time for the household sample and 18.5% for the individual sample. See Mitchell *et al.* (1982 pp. 20-21).

REFERENCES

- BATEMAN, F., and FOUST, J.D. (1974). A sample of rural households selected from the 1860 manuscript censuses. *Agricultural History*, 48, 75-93.
- BEBBINGTON, A.D. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 135.
- BODE, F.A., and GINTER, D.E. (1984). A critique of land holding variables in the 1860 census and the Parker-Gallman sample. *Journal of Interdisciplinary History*, 15, 277-295.
- DARROCH, A.G., and ORNSTEIN, M.D. (1980). Ethnicity and occupational structure in Canada in 1871: the vertical mosaic in historical perspective. *Canadian Historical Review*, 61, 305-333.
- FOGEL, R.W., and ENGERMAN, S.L. (1974). *Time on the Cross: Evidence and Methods*. Boston: Little, Brown and Co.
- FOUST, J.D. (1975). *The Yeoman Farmer and Westward Expansion of U.S. Cotton Production*. New York: Arno Press.
- GRAHAM, S.N. (1980). *1900 Public Use Sample: User's Handbook*. Seattle: Centre for Studies in Demography and Ecology, University of Washington.
- HAMMARBERG, M.A. (1971). Designing a sample from incomplete historical lists. *American Quarterly*, 23, 542-561.
- HAMMARBERG, M.A. (1977). A sampling design for Mormon Utah, 1880. *Journal of Interdisciplinary History*, 7, 453-476.
- HOLT, D., SCOTT, A.J., and EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Series A*, 143, 303-320.
- JOHNSON, R.C. (1978a). A procedure for sampling manuscript census schedules. *Journal of Interdisciplinary History*, 8, 513-530.
- JOHNSON, R.C. (1978b). The 1900 census sampling project: methods and procedures for sampling and data entry. *Historical Methods*, 11, 147-151.
- McINNIS, R.M. (1977). Childbearing and land availability: some evidence from individual household data. *Population Patterns in the Past*, (R.D. Lee ed.), New York: Academic Press, 201-227.

- MITCHEL, S.P., LINK, D.G., and HANIS, E.H. (1982). *Final Report: Determination of Procedures and Costs for the Production of a Machine Readable Edition of the 1881 Census of Canada*. DSS Contract Ser. No. 0SU80-00326.
- ORNSTEIN, M.D. (1978). The design of a sample of households from the 1871 census of Canada. Unpublished manuscript, York University, Toronto.
- ORNSTEIN, M.D., and DARROCH, G.O. (1978). National mobility studies in past time: a sample strategy. *Historical Methods*, 11, 152-161.
- RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex surveys: chi-square tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- SCOTT, A.J., and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.
- SMITH, D.S. (1978). A community-based sample of the older population from the 1880 and 1900 United States manuscript census. *Historical Methods*, 11, 67-74.
- SOLTOW, L. (1975). *Men and Wealth in the United States 1850-1870*. New Haven: Yale University Press.
- STATISTICS CANADA (1975). *1971 Census of Canada: Public Use Sample Tapes User Documentation*, Ottawa.
- STATISTICS CANADA (1979). *1976 Census of Canada: Public Use Sample Tapes User Documentation*, Ottawa.
- SWEIRENGA, R.P. (1983). Quantitative methods in rural landholding. *Journal of Interdisciplinary History*, 13, 787-808.

Estimation of Total for Two Characters in Multiple Frame Surveys

B.C. SAXENA, P. NARAIN, and A.K. SRIVASTAVA¹

ABSTRACT

In this paper estimation of multiple characters in multiple frame surveys has been investigated. The gain due to two character study in a common survey, over separate surveys for individual characters, has been obtained. Cost comparison is also made between two character multi frame survey and two character single frame survey.

KEY WORDS: Multi-character survey; Post-stratified estimate; Optimization; Cost comparison.

1. INTRODUCTION

The technique of multiple frame surveys was suggested by Hartley (1962) and subsequently discussed by Lund (1968), Hartley (1974), Vogel (1975), Armstrong (1979), etc. Lund suggested an alternate to Hartley's estimator utilizing the actual division in the sample among various domains. Hartley (1974) further considered the problem with more general approach applicable to various sampling designs. He observed that most potential multiple frame situations employed different types of units in their respective frames. Bosecker and Ford (1976) extended Hartley's estimator to take advantage of stratification within the overlap domain. Serrurier and Phillips (1976) and Armstrong (1978) tested multiple frame techniques in agricultural surveys. The utility of multiple frame survey has been demonstrated in a wide variety of situations. In sample surveys, sometimes interest lies not only in the estimation of single character but several characters are required to be studied simultaneously. For a proper utilization of resources this is often achieved through integrated surveys. For instance, for estimating the production of vegetable crops, a single survey is planned to estimate the production of several vegetable crops. Also, besides the frame of all vegetable growers, another incomplete but relatively easily accessible frame of important vegetable growers may be utilized. In this paper, the estimation of total for two characters in multiple frame surveys has been considered. The advantage of studying more than one character in a single survey over the situation when independent surveys are planned for individual characters in a multiple frame situation, is also investigated.

2. ESTIMATOR

Let there be two overlapping frames A and B of sizes N_A and N_B respectively. In multiple frame surveys two samples of sizes n_A and n_B are selected independently by simple random sampling from frames A and B respectively. The overlapping frames generate domains a , b and ab defined as follows:

- a : Consisting of units belonging to frame A only,
- b : Consisting of units belonging to frame B only,
- ab : Units belonging to both A and B frames.

¹ B.C. Saxena, P. Narain, and A.K. Srivastava, Indian Agricultural Statistics Research Institute, New Delhi, India.

The sample sizes n_A and n_B are split into sizes n_a , n_{ab} and n_b , n_{ba} such that n_a and n_{ab} are the number of units out of n_A units belonging to domains a and ab respectively. Similarly n_b and n_{ba} are the split of n_B units belonging to domains b and ab respectively. In the multi-character study, there will be further split of these domains generating sub-domains as follows:

Let there be two characters $y_{(1)}$ and $y_{(2)}$ under study. Then each of the usual domains a , ab and b are further subdivided as $a(1)$, $a(12)$, $a(2)$, $ab(1)$, $ab(12)$, $ab(2)$ and $b(1)$, $b(12)$, $b(2)$ respectively. Here, $a(1)$, $a(12)$ and $a(2)$ are the sub-domains consisting of units having character $y_{(1)}$, both $y_{(1)}$ and $y_{(2)}$, and $y_{(2)}$ only respectively in domain a . Similar explanation holds for other sub-domains $ab(1)$, $ab(12)$ etc. Thus the sample split in two character study will be as follows:

$$n_A = n_a + n_{ab}$$

where

$$n_a = n_{a(1)} + n_{a(2)} + n_{a(12)} \quad \text{and} \quad n_{ab} = n_{ab(1)} + n_{ab(2)} + n_{ab(12)},$$

and

$$n_B = n_b + n_{ba}$$

where

$$n_b = n_{b(1)} + n_{b(2)} + n_{b(12)} \quad \text{and} \quad n_{ba} = n_{ba(1)} + n_{ba(2)} + n_{ba(12)}.$$

Here $n_{a(1)}$, $n_{a(2)}$, etc. are the split of n_a units belonging to sub-domains $a(1)$, $a(2)$, etc. If we confine to one character then define

$$n_{A(1)} = n_{a(1)} + n_{a(12)} + n_{ab(1)} + n_{ab(12)},$$

$$n_{B(1)} = n_{b(1)} + n_{b(12)} + n_{ba(1)} + n_{ba(12)}.$$

Similarly, for the second character, $n_{A(2)}$ and $n_{B(2)}$ are defined. The estimate of the total for the first character is given by

$$\begin{aligned} \hat{Y}^{(1)} = & \hat{Y}_{a(1)} + \hat{Y}_{a(12)}^{(1)} + p_1 \hat{Y}_{ab(1)} + q_1 \hat{Y}_{ba(1)} + p_2 \hat{Y}_{ab(12)}^{(1)} + \\ & + q_2 \hat{Y}_{ba(12)}^{(1)} + \hat{Y}_{b(1)} + \hat{Y}_{b(12)}^{(1)} \end{aligned} \quad (1)$$

where $\hat{Y}_{a(1)}$, $\hat{Y}_{a(12)}^{(1)}$, etc. are the estimated totals for character $y_{(1)}$ of the respective sub-domains. In the subsequent discussion, for the domains in which both the characters are available, the super script corresponds to the character under consideration. For the domains having only one character the super script is not used since the domain evidently corresponds to the character.

Also, $p_1 + q_1 = 1$ and $p_2 + q_2 = 1$. Define $\bar{y}_{a(1)}$, $\bar{y}_{a(2)}$, etc. as the sample means for respective sub-domains for character $y_{(1)}$ and $y_{(2)}$ respectively.

Thus,

$$\begin{aligned}\hat{Y}^{(1)} = & N_{a(1)}\bar{y}_{a(1)} + N_{a(12)}\bar{y}_{a(12)}^{(1)} + N_{ab(1)}(p_1\bar{y}_{ab(1)} + q_1\bar{y}_{ba(1)}) \\ & + N_{ab(12)}(p_2\bar{y}_{ab(12)}^{(1)} + q_2\bar{y}_{ba(12)}^{(1)}) \\ & + N_{b(12)}\bar{y}_{b(12)}^{(1)} + N_{b(1)}\bar{y}_{b(1)}.\end{aligned}\quad (2)$$

Similarly for the second character, we have

$$\begin{aligned}\hat{Y}^{(2)} = & N_{a(2)}\bar{y}_{a(2)} + N_{a(12)}\bar{y}_{a(12)}^{(2)} + N_{ab(2)}(p_3\bar{y}_{ab(2)} + q_3\bar{y}_{ba(2)}) \\ & + N_{ab(12)}(p_4\bar{y}_{ab(12)}^{(2)} + q_4\bar{y}_{ba(12)}^{(2)}) + N_{b(12)}\bar{y}_{b(12)}^{(2)} \\ & + N_{b(2)}\bar{y}_{b(2)}\end{aligned}\quad (3)$$

where

$$p_3 + q_3 = 1 \text{ and } p_4 + q_4 = 1.$$

2.1 Variance of the Estimator

The conditional variance of the post-stratified estimates $\hat{Y}^{(1)}, \hat{Y}^{(2)}$ for given sub-domain sample sizes ignoring the finite population correction may be written as

$$\begin{aligned}V(\hat{Y}^{(1)} | n_{a(1)}, n_{a(12)}, \text{ etc.}) = & N_{a(1)}^2 \frac{\sigma_{a(1)}^2}{n_{a(1)}} + N_{a(12)}^2 \frac{\sigma_{a(12)}^{(1)2}}{n_{a(12)}} \\ & + N_{ab(1)}^2 \left(p_1^2 \frac{\sigma_{ab(1)}^2}{n_{ab(1)}} + q_1^2 \frac{\sigma_{ba(1)}^2}{n_{ba(1)}} \right) \\ & + N_{ab(12)}^2 \left(p_2^2 \frac{\sigma_{ab(12)}^{(1)2}}{n_{ab(12)}} + q_2^2 \frac{\sigma_{ba(12)}^{(1)2}}{n_{ba(12)}} \right) + N_{b(1)}^2 \frac{\sigma_{b(1)}^2}{n_{b(1)}} \\ & + N_{b(12)}^2 \frac{\sigma_{b(12)}^{(1)2}}{n_{b(12)}}\end{aligned}\quad (4)$$

The unconditional variance of $\hat{Y}^{(1)}$ is approximately given by

$$\begin{aligned}V(\hat{Y}^{(1)}) = & \frac{N_A}{n_A} \left\{ N_{a(1)}\sigma_{a(1)}^2 + N_{a(12)}\sigma_{a(12)}^{(1)2} + p_1^2 N_{ab(1)}\sigma_{ab(1)}^2 \right. \\ & + p_2^2 N_{ab(12)}\sigma_{ab(12)}^{(1)2} \left. \right\} + \frac{N_B}{n_B} \left\{ N_{b(1)}\sigma_{b(1)}^2 + N_{b(12)}\sigma_{b(12)}^{(1)2} \right. \\ & + q_1^2 N_{ab(1)}\sigma_{ab(1)}^2 + q_2^2 N_{ab(12)}\sigma_{ab(12)}^{(1)2} \left. \right\}\end{aligned}\quad (5)$$

which is equal to the variance for stratified sampling with proportional allocation.

Similarly,

$$\begin{aligned}
 V(\hat{Y}^{(2)}) = & \frac{N_A}{n_A} \left\{ (N_{a(2)} \sigma_{a(2)}^2 + N_{a(12)} \sigma_{a(12)}^{(2)^2} + p_3^2 N_{ab(2)} \sigma_{ab(2)}^2 \right. \\
 & + p_4^2 N_{ab(12)} \sigma_{ab(12)}^{(2)^2} \left. \right\} + \frac{N_B}{n_B} \left\{ N_{b(2)} \sigma_{b(2)}^2 + N_{b(12)} \sigma_{b(12)}^{(2)^2} \right. \\
 & + q_3^2 N_{ab(2)} \sigma_{ab(2)}^2 + q_4^2 N_{ab(12)} \sigma_{ab(12)}^{(2)^2} \left. \right\} \quad (6)
 \end{aligned}$$

where $\sigma_{a(1)}^2$, $\sigma_{a(2)}^2$, etc. are the variances for the two characters in the respective sub-domains.

For optimization of p_i 's ($i = 1, 2, 3, 4$) for a common survey a combination of individual variances needs to be minimized subject to the fixed total cost for the combined survey. Consider the simplest linear combination

$$F = V(\hat{Y}^{(1)}) + V(\hat{Y}^{(2)}).$$

For the common survey, a suitable cost function may be considered as follows:

$$\begin{aligned}
 C' = & C_1(n_{a(1)} + n_{ab(1)}) + C_2(n_{a(12)} + n_{ab(12)}) + C_3(n_{a(2)} + n_{ab(2)}) \\
 & + C_4(n_{b(1)} + n_{ba(1)}) + C_5(n_{b(12)} + n_{ba(12)}) + C_6(n_{b(2)} + n_{ba(2)}) \quad (7)
 \end{aligned}$$

where C_1 is the cost per unit in sub-domain $a(1)$, $ab(1)$; C_2 in $a(12)$, $ab(12)$; C_3 in $a(2)$, $ab(2)$ of frame A . Similarly C_4 , C_5 and C_6 are the cost per unit from frame B . In the above cost function random sample sizes are involved. Consider the expected cost

$$C = E(C') = n_A(C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3) + n_B(C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6) \quad (8)$$

where

$$\begin{aligned}
 \Phi_1 &= \frac{N_{a(1)} + N_{ab(1)}}{N_A}, \quad \Phi_2 = \frac{N_{a(12)} + N_{ab(12)}}{N_A}, \\
 \Phi_3 &= \frac{N_{a(2)} + N_{ab(2)}}{N_A}, \quad \Phi_4 = \frac{N_{b(1)} + N_{ba(1)}}{N_B}, \\
 \Phi_5 &= \frac{N_{b(12)} + N_{ba(12)}}{N_B}, \quad \Phi_6 = \frac{N_{b(2)} + N_{ba(2)}}{N_B}.
 \end{aligned}$$

Or

$$C = n_A C_A + n_B C_B \quad (9)$$

where

$$C_A = C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3 \quad \text{and} \quad C_B = C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6.$$

In order to get the optimum p_i 's as also n_A and n_B , the function F is to be minimised subject to the expected cost function as given in (9). The weight variables p_i 's and sample sizes are obtained as follow using Lagrange multiplier:

$$\frac{P_1}{q_1} = \frac{P_2}{q_2} = \frac{P_3}{q_3} = \frac{P_4}{q_4} = \frac{N_B n_A}{n_B N_A} = \frac{P}{q} \text{ (say),} \quad (10)$$

and

$$\begin{aligned} \frac{n_A^2}{N_A} &= \gamma \frac{K_5 + K_1 p_1^2 + K_2 p_2^2 + K_3 p_3^2 + K_4 p_4^2}{C_A}, \\ \frac{n_B^2}{N_B} &= \gamma \frac{K_6 + K_1 q_1^2 + K_2 q_2^2 + K_3 q_3^2 + K_4 q_4^2}{C_B}, \end{aligned} \quad (11)$$

with γ determined to meet the expected cost and

$$\begin{aligned} K_1 &= N_{ab(1)} \sigma_{ab(1)}^2, \quad K_2 = N_{ab(12)} \sigma_{ab(12)}^{(1)2}, \\ K_3 &= N_{ab(2)} \sigma_{ab(2)}^2, \quad K_4 = N_{ab(12)} \sigma_{ab(12)}^{(2)2}, \\ K_5 &= N_{a(1)} \sigma_{a(1)}^2 + N_{a(2)} \sigma_{a(2)}^2 + N_{a(12)} (\sigma_{a(12)}^{(1)2} + \sigma_{a(12)}^{(2)2}), \\ K_6 &= N_{b(1)} \sigma_{b(1)}^2 + N_{b(2)} \sigma_{b(2)}^2 + N_{b(12)} (\sigma_{b(12)}^{(1)2} + \sigma_{b(12)}^{(2)2}). \end{aligned} \quad (12)$$

From (10) and (11), we get

$$\frac{q^2 N_B C_B}{p^2 N_A C_A} = \frac{K_6 + (K_1 + K_2 + K_3 + K_4) q^2}{K_5 + (K_1 + K_2 + K_3 + K_4) p^2} \quad (13)$$

This is a bi-quadratic in p and can be solved for p . The optimum sampling fractions can be obtained from (11). A practical case commonly met in multiple frame situations is when one of the frames has got 100% coverage. Consider 100% coverage by the frame A then $N_{b(1)} = N_{b(2)} = N_{b(12)} = 0$.

In this case (13) reduces to

$$p^2 = \frac{\alpha}{q - \alpha} \frac{K_5}{K_1 + K_2 + K_3 + K_4} \quad (14)$$

where

$$q = \frac{C_A}{C_B} \quad \text{and} \quad \alpha = \frac{N_B}{N_A}.$$

Assume that

$$\sigma_{a(1)}^2 = \sigma_{a(12)}^{(1)2}, \sigma_{a(2)}^2 = \sigma_{a(12)}^{(2)2}, \sigma_{ab(1)}^2 = \sigma_{ab(12)}^{(1)2}, \sigma_{ab(2)}^2 = \sigma_{ab(12)}^{(2)2}. \quad (15)$$

These assumptions appear plausible since the variability of one character is not likely to be affected by the presence or absence of the other character. Then p^2 reduces to

$$p^2 = \frac{\alpha}{\varrho - \alpha} \left\{ \frac{\sigma_{a(1)}^2(N_{a(1)} + N_{a(12)}) + \sigma_{a(2)}^2(N_{a(2)} + N_{a(12)})}{\sigma_{ab(1)}^2(N_{ab(1)} + N_{ab(12)}) + \sigma_{ab(2)}^2(N_{ab(2)} + N_{ab(12)})} \right\}$$

or

$$p^2 = \frac{(1 - \alpha)\Phi'_2}{(\varrho - \alpha)} \left\{ \frac{\Phi'_3(\xi_1 + \xi_2) + (1 - \xi_1)}{\Phi'_4(\xi_3 + \xi_4) + (1 - \xi_3)} \right\} \quad (16)$$

where

$$\Phi'_1 = \frac{\sigma_{a(1)}^2}{\sigma_{ab(1)}^2}, \Phi'_2 = \frac{\sigma_{a(2)}^2}{\sigma_{ab(2)}^2}, \Phi'_3 = \frac{\sigma_{a(1)}^2}{\sigma_{a(2)}^2}, \Phi'_4 = \frac{\sigma_{ab(1)}^2}{\sigma_{ab(2)}^2}$$

and

$$\xi_1 = \frac{N_{a(1)}}{N_a}, \xi_2 = \frac{N_{a(12)}}{N_a}, \xi_3 = \frac{N_{ab(1)}}{N_{ab}}, \xi_4 = \frac{N_{ab(12)}}{N_{ab}}.$$

Using that $N_{ab} = N_B$, $N_{a(2)} + N_{a(12)} = N_a - N_{a(1)}$ and $N_{ab(2)} + N_{ab(12)} = N_{ab} - N_{ab(1)}$, it may be seen that the above expression of p^2 reduces to the usual form in uni-character case since $\xi_1 = \xi_3 = 1$ and $\xi_2 = \xi_4 = 0$. It may be remarked that the domain variances are generally not known as such these values are based either on prior knowledge or some guessed values. The optimality of p^2 is effected to that extent.

3. COMPARISON OF MULTI-CHARACTER SURVEY WITH INDEPENDENT UNI-CHARACTER SURVEYS IN MULTIPLE FRAME SITUATIONS

Multi-character surveys are planned with a view to economise the available resources and it is expected that a common survey is likely to score over independent uni-character surveys taking into account the cost and efficiency. In this situation the extent of gain due to a common multiple frame survey is investigated.

In a single character study for character $y_{(1)}$ (say), consider simple random samples of sizes n_A and n_B from the frames A and B respectively. Here we assume that the only frames used before are available, not the reduced frame for each character. Define N_{A1} , N_{B1} , n_{A1}^* , and n_{B1}^* as the population sizes and sample sizes respectively with character $y_{(1)}$. Here,

n_{A1}^* and n_{B1}^* are the random sample sizes with $E(n_{A1}^*) = n_A N_{A1}/N_A$ and $E(n_{B1}^*) = n_B N_{B1}/N_B$. In this case, the estimator $\hat{Y}^{(1)*}$ and its variance are as follows:

$$\begin{aligned}\hat{Y}^{(1)*} &= (N_{a(1)} + N_{a(12)})\bar{y}_{(a(1), a(12))} \\ &+ (N_{ab(1)} + N_{ab(12)})(p'\bar{y}_{(ab(1), ab(12))} + q'\bar{y}_{(ba(1), ba(12))}) \\ &+ (N_{b(1)} + N_{b(12)})\bar{y}_{(b(1), b(12))}\end{aligned}$$

where p' , q' are weight variables such that $p' + q' = 1$ and $\bar{y}_{(a(1), a(12))}$, $\bar{y}_{(ab(1), ab(12))}$, etc. are sample means for the sample from combined respective domains, e.g. $\bar{y}_{(a(1), a(12))}$ is the mean of sample units coming from domain $a(1)$ and $a(12)$.

$$\begin{aligned}V(\hat{Y}^{(1)*}) &= \frac{N_A}{n_A}(N_{a(1)}\sigma_{a(1)}^2 + N_{a(12)}\sigma_{a(12)}^2) \\ &+ (p'^2\frac{N_A}{n_A} + q'^2\frac{N_B}{n_B})(N_{ab(1)}\sigma_{ab(1)}^2 + N_{ab(12)}\sigma_{ab(12)}^2) \\ &+ \frac{N_B}{n_B}(N_{b(1)}\sigma_{b(1)}^2 + N_{b(12)}\sigma_{b(12)}^2).\end{aligned}\quad (17)$$

In this case, the cost function is of the form

$$C = C_1 n_{A1}^* + C_4 n_{B1}^*$$

and expected cost is given by C^* as

$$C^* = C_1 \frac{n_A}{N_A} N_{A1} + C_4 \frac{n_B}{N_B} N_{B1} = C'_A n_A + C'_B n_B \quad (18)$$

where $C'_A = C_1 N_{A1}/N_A$ and $C'_B = C_4 N_{B1}/N_B$.

For simplicity, we assume 100% coverage by frame A , equality of variances as in (15), and $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$. Based on these assumptions, the cost C^* with n_A and n_B which minimize the variance (17) is given by (see Appendix for derivation).

$$C^* = \frac{(\xi_1 + \xi_2)[\{C_1(1 + \alpha_1^*)(\Phi_1' + \alpha_1^* p'^2)\}^{1/2} + \alpha_1^*(C_4 q'^2)^{1/2}]^2}{1 - \alpha \left\{ \frac{(\Phi_1' + \alpha_1^* p^2)}{n_A} + \frac{\alpha \alpha_1^* q^2}{n_B} \right\}}$$

where

$$\alpha_1^* = \frac{\alpha}{1 - \alpha} \frac{\xi_3 + \xi_4}{\xi_1 + \xi_2}.$$

Similarly, for the separate survey for the 2nd character, the cost is obtained as

$$C^{**} = \frac{(1 - \xi_1) \left[\left\{ C_3(1 + \alpha_2^*)(\Phi_2' + \alpha_2^*p^{n_2}) \right\}^{1/2} + \alpha_2^*(C_6q^{n_2})^{1/2} \right]^2}{\frac{1}{1 - \alpha} \left\{ \frac{(\Phi_2' + \alpha_2^*p^2)}{n_A} + \frac{\alpha\alpha_2^*q^2}{n_B} \right\}} \quad (19)$$

where

$$p^{n_2} = \frac{K\Phi_2'}{1 + \alpha_2^*(1 - K)}, \quad \alpha_2^* = \frac{\alpha}{1 - \alpha} \frac{1 - \xi_3}{1 - \xi_1}.$$

For the combined character study, the total cost C for 100% coverage by the frame A is given by (8).

Thus

$$C = \frac{n_A}{N_A} [C_1(N_{a(1)} + N_{ab(1)}) + C_2(N_{a(12)} + N_{ab(12)}) + C_3(N_{a(2)} + N_{ab(2)})] \\ + \frac{n_B}{N_B} [C_4N_{ab(1)} + C_5N_{ab(12)} + C_6N_{ab(2)}].$$

Using assumptions in costs (i.e. $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$) we get

$$C = C_2n_A \left\{ (1 - \alpha) \left\{ \varrho_1\xi_1 + \xi_2 + \varrho_3(1 - \xi_1 - \xi_2) \right\} + \right. \\ \left. \alpha \left\{ \varrho_1\xi_3 + \xi_4 + \varrho_3(1 - \xi_3 - \xi_4) \right\} + \frac{K}{r} \left\{ \varrho_1\xi_3 + \xi_4 + \varrho_3(1 - \xi_3 - \xi_4) \right\} \right\} \quad (20)$$

where $r = n_A/n_B$, $\varrho_1 = C_1/C_2$ and $\varrho_3 = C_3/C_2$.

But in combined character study (n_A/n_B) Opt. = $p/\alpha q$ where p is given by (16). Thus the gain may be obtained from the ratio.

$$\frac{C^* + C^{**}}{C} = \frac{\frac{(\xi_1 + \xi_2)\varrho_1T_1^2}{(\Phi_1' + \alpha_1^*p)} + \frac{(1 - \xi_1)\varrho_3T_2^2}{(\Phi_3' + \alpha_3^*p)}}{\frac{\left\{ \varrho_1\xi_1 + \xi_2 + \varrho_3(1 - \xi_1 - \xi_2) \right\} + \left\{ \varrho_1\xi_3 + \xi_4 + \varrho_3(1 - \xi_3 - \xi_4) \right\} \left\{ \frac{r\alpha + K}{r(1 - \alpha)} \right\}}}{\quad} \quad (21)$$

where

$$T_1 = \left\{ (\Phi'_1 + \alpha_1^* p'^2)(1 + \alpha_1^*) \right\}^{1/2} + \alpha_1^* q' \sqrt{K}$$

$$T_2 = \left\{ (\Phi'_2 + \alpha_2^* p''^2)(1 + \alpha_2^*) \right\}^{1/2} + \alpha_2^* q'' \sqrt{K}.$$

K can be determined as follows: Using the definitions of C_A , C_B , Φ_i 's ($i = 1, \dots, 6$) and equation (A.1), we obtain

$$\frac{C_A}{C_B} = \frac{1}{K} \frac{\varrho_1 \Phi_1 + \Phi_2 + \varrho_3 \Phi_3}{\varrho_1 \Phi_4 + \Phi_5 + \varrho_3 \Phi_6} = \varrho,$$

and thus

$$K = \varrho^{-1} \left\{ \alpha + (1 - \alpha) \frac{\varrho_1 \xi_1 + \xi_2 + \varrho_3(1 - \xi_1 - \xi_2)}{\varrho_1 \xi_3 + \xi_4 + \varrho_3(1 - \xi_3 - \xi_4)} \right\}. \quad (22)$$

The expression in (21) may be used to obtain the gain in cost due to studying both the character simultaneously in comparison to independent individual surveys. The percent gain G is thus given by

$$G = \left(\frac{C^* + C^{**}}{C} - 1 \right) \times 100$$

In the above cost comparison, the expected costs, C , C^* and C^{**} do not include the overhead costs for the combined or individual surveys, however, it is expected that the sum of overhead costs pertaining to individual surveys would be much larger than the corresponding overhead cost for the combined survey. Therefore, the actual gain in costs due to common multiple frame surveys compared to independent surveys will be larger than the percent gain G defined above.

The expression (21) reduced substantially under the assumptions $\Phi'_1 = \Phi'_2 = \Phi$ (say) and $\xi_1 = \xi_2 = \xi_3 = \xi_4 = \xi$ (say).

From (22) $\varrho = 1/K$ and from (16) since $\Phi'_1/\Phi'_2 = \Phi'_3/\Phi'_4$, the p^2 reduces as follows:

$$p^2 = \frac{K(1 - \alpha)}{1 - k\alpha} \Phi.$$

Also $\alpha_1^* = \alpha_2^* = \alpha/(1 - \alpha)$.

Therefore, from (A.1)

$$p' = p'' = \left\{ \frac{K(1 - \alpha)\Phi}{1 - K\alpha} \right\}^{1/2}.$$

Thus

$$T_1 = T_2 = \left\{ \Phi \frac{1 - K\alpha}{1 - \alpha} \right\}^{1/2} + \frac{\alpha \sqrt{K}}{1 - \alpha}.$$

With all these substitutions in (21) $(C^* + C^{**})/C$ simplifies as follows:

$$\begin{aligned}\frac{C^* + C^{**}}{C} &= \frac{T_1^2}{\Phi + \alpha_1^* p} \times \frac{2\xi q_1 + (1 - \xi)q_3}{\left\{ \frac{r\alpha + K}{K(1 - \alpha)} + 1 \right\} \{ q_1\xi + \xi + q_3(1 - 2\xi) \}} \\ &= \frac{T_1^2}{(\Phi + \alpha_1^* p) \left\{ \frac{r\alpha + K}{K(1 - \alpha)} + 1 \right\}} \times \frac{q_3 + \xi(2q_1 - q_3)}{q_3 + \xi(q_1 + 1 - 2q_3)} \\ &= \frac{q_3 + \xi(2q_1 - q_3)}{q_3 + \xi(1 + q_1 - 2q_3)}\end{aligned}$$

where $r = (n_A/n_B)$ opt. = $p/\alpha q$ from (10).

Hence,

$$G = \frac{\xi(q_1 + q_3 - 1)}{q_3 + \xi(1 + q_1 - 2q_3)} \times 100.$$

The equality of ξ_i 's does not seem to be realistic assumption. The value of G , has therefore been calculated using (19) for realistic and representative combinations of parameters and are presented in Table. 1.

This table indicates that there is a definite gain due to integration of multiple frame surveys for both the characters in comparison to separate individual surveys. The gain increases with increasing values of q_1 and q_3 .

4. COMPARISON OF TWO CHARACTER MULTIPLE FRAME SURVEYS WITH SINGLE FRAME SURVEY

Comparison of two frame survey with single frame surveys for study of two characters is of practical interest. For single character a similar study was carried out by Hartley (1962). On similar lines the relative reduction in cost was obtained as

$$R = \left(1 + \frac{\alpha q}{p q} \right)^2 \bigg/ \left\{ 1 + \frac{\alpha q(1 + p)}{p^2 q} \right\}$$

where p^2 is given by (16), $q = C_A/C_B$ and $\alpha = N_{ab}/N_A$

The reduction in cost due to multiple frame over a single frame survey is tabulated in Table 2 for some set of parametric values. The table indicates considerable cost reduction.

Table 1
Percent Gain in Cost for Common multiple Frame Survey
for Both Characters over Individual Surveys,
When $q = 10$, $\Phi'_1 = 0.25$, $\Phi'_2 = 0.5$, $\Phi'_3 = 1$, $\alpha = 0.5$.

q_1	q_3					
	0.3	0.4	0.5	0.6	0.7	0.8
	$\xi_1 = 0.2, \xi_2 = 0.2, \xi_3 = 0.4, \xi_4 = 0.2$					
0.3						1.5
0.4					1.7	4.2
0.5				1.8	4.4	6.7
0.6			1.8	4.5	6.9	8.9
0.7		1.7	4.6	7.0	9.1	10.9
0.8	1.7	4.6	7.1	9.3	11.2	12.8
0.9	4.5	7.1	9.4	11.3	13.0	14.5
	$\xi_1 = 0.2, \xi_2 = 0.4, \xi_3 = 0.2, \xi_4 = 0.4$					
0.3						4.5
0.4					4.6	9.3
0.5				4.8	9.6	14.0
0.6			4.9	9.9	14.3	18.3
0.7		5.1	10.1	14.7	18.8	22.5
0.8	5.2	10.4	15.1	19.3	23.1	26.5
0.9	10.8	15.5	19.8	23.6	27.1	30.3

Table 2
Reduction in Cost for Constant Variances
When $\Phi'_1 = 0.25$, $\Phi'_2 = 0.5$, $\Phi'_3 = 1$, and $\xi_1 = 0.2$, $\xi_2 = 0.3$, $\xi_4 = 0.4$.

q	α					
	0.5	0.6	0.7	0.8	0.9	0.95
100	.227	.175	.132	.094	.059	.040
20	.304	.254	.200	.169	.127	.101
10	.367	.321	.279	.238	.193	.164
5	.462	.423	.387	.351	.308	.277
2	.661	.646	.634	.621	.599	.578
1	.876	.895	.918	.943	.971	.985

APPENDIX

Minimizing the variance (17) with respect to C^* with the assumption of 100% coverage by frame A and the equality of variances, the optimum solution for p' is obtained as

$$p'^2 = \frac{1 - \alpha}{\varrho' - \alpha} \left\{ \frac{\sigma_{a(1)}^2(\xi_1 + \xi_2)}{\sigma_{ab(1)}^2(\xi_3 + \xi_4)} \right\}$$

with

$$\varrho' = \frac{C'_A}{C'_B}.$$

Using $N_{A1} = N_{a(1)} + N_{a(12)} + N_{ab(1)} + N_{ab(12)}$ and $N_{B1} = N_{ab(1)} + N_{ab(12)}$, ϱ' can be written as

$$\begin{aligned} \varrho' &= \frac{C_1}{C_4} \alpha \frac{N_a(\xi_1 + \xi_2) + N_{ab}(\xi_3 + \xi_4)}{N_{ab}(\xi_3 + \xi_4)} \\ &= \frac{C_1}{C_4} \alpha \left(\frac{1 - \alpha}{\alpha} \frac{\xi_1 + \xi_2}{\xi_3 + \xi_4} + 1 \right) \\ &= \frac{\alpha}{K} \left(\frac{1}{\alpha_1^*} + 1 \right) \end{aligned}$$

where

$$\alpha_1^* = \frac{\alpha}{1 - \alpha} \frac{\xi_3 + \xi_4}{\xi_1 + \xi_2}.$$

Then we have

$$\begin{aligned} p'^2 &= \frac{1 - \alpha}{\frac{\alpha}{K} \left(\frac{1}{\alpha_1^*} + 1 \right) - \alpha} \frac{\xi_1 + \xi_2}{\xi_3 + \xi_4} \Phi_1' \\ &= \frac{K \Phi_1'}{1 + \alpha_1^*(1 - K)} \end{aligned} \tag{A.1}$$

Define

$$\lambda_1 = (N_{a(1)} + N_{a(12)})\sigma_{a(1)}^2 + p^2(N_{ab(1)} + N_{ab(12)})\sigma_{ab(1)}^2,$$

$$\lambda_2 = q^2(N_{ab(1)} + N_{ab(12)})\sigma_{ab(1)}^2,$$

$$\lambda_3 = (N_{a(1)} + N_{a(12)})\sigma_{a(1)}^2 + p'^2(N_{ab(1)} + N_{ab(12)})\sigma_{ab(1)}^2,$$

$$\lambda_4 = q'^2(N_{ab(1)} + N_{ab(12)})\sigma_{ab(1)}^2.$$

With the p' in (A.1), the optimum sample sizes will be

$$\begin{aligned} \frac{n_{AO}^2}{N_A} &= \gamma' \frac{(N_{a(1)} + N_{a(12)})\sigma_{a(1)}^2 + p'^2(N_{ab(1)} + N_{ab(12)})\sigma_{ab(1)}^2}{C'_A} \\ &= \gamma' \frac{\lambda_3}{C'_A} \end{aligned}$$

$$\frac{n_{BO}^2}{N_B} = \gamma' \frac{q'^2(N_{ab(1)} + N_{ab(12)})\sigma_{ab(1)}^2}{C'_B} = \gamma' \frac{\lambda_4}{C'_B}$$

with γ' determined with respect to (18). From this we get

$$\frac{n_{BO}}{n_{AO}} = \frac{N_B}{N_A} \left(\frac{C_1 N_{A1} \lambda_4}{C_4 N_{B1} \lambda_3} \right)^{1/2}. \tag{A.2}$$

Also, the variances given by (5) and (17) at optimum sample sizes can be written as

$$V(\hat{Y}^{(1)}) = \frac{N_A}{n_A} \lambda_1 + \frac{N_B}{n_B} \lambda_2 \tag{A.3}$$

$$V(\hat{Y}^{*(1)}) = \frac{N_A}{n_{AO}} \lambda_3 + \frac{N_B}{n_{BO}} \lambda_4.$$

Equating the above variances and using (A.2), we obtain expression for n_{AO} and n_{BO} in terms of n_A and n_B as follows:

$$\frac{n_{AO}}{N_A} = \frac{\lambda_3 + \left(\frac{C_4 \lambda_3 \lambda_4 N_{B1}}{C_1 N_{A1}} \right)^{1/2}}{\frac{N_A}{n_A} \lambda_1 + \frac{N_B}{n_B} \lambda_2}$$

and

$$\frac{n_{BO}}{N_B} = \frac{\lambda_4 + \left(\frac{C_1 \lambda_3 \lambda_4 N_{A1}}{C_4 N_{B1}} \right)^{1/2}}{\frac{N_A}{n_A} \lambda_1 + \frac{N_B}{n_B} \lambda_2}.$$

Using these relationships, the cost C^* may be obtained as

$$C^* = \frac{(\xi_1 + \xi_2) \left[\left\{ C_1 (1 + \alpha_1^*) (\Phi_1' + \alpha_1^* p'^2) \right\}^{1/2} + \alpha_1^* (C_4 q'^2)^{1/2} \right]^2}{1 - \alpha \left\{ \frac{(\Phi_1' + \alpha_1^* p^2)}{n_A} + \frac{\alpha \alpha_1^* q^2}{n_B} \right\}} \quad (\text{A.4})$$

REFERENCES

- ARMSTRONG, B. (1979). Test for multiple frames sampling technique for agricultural survey: New Brunswick, 1978. *Survey Methodology*, 5, 178-199.
- BOSECKER, R.R., and FORD, B.L. (1976). Multiple frame estimation with stratified overlap domain. *American Statistical Association Proceedings of the Social Statistics Section*, 219-224.
- HARTLEY, H.O. (1962). Multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 203-206.
- HARTLEY, H.O. (1974). Multiple frame methodology and selected application. *Sankhya*, Series C, 36, 99-118.
- LUND, R.E. (1968). Estimation in multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 282-288.
- SERRURIER, D., and PHILLIPS, J. (1976). Double frame Ontario pilot hog surveys. *Survey Methodology*, 2, 138-170.
- VOGEL, F.A. (1975). Surveys with overlapping frames, problems in application. *American Statistical Association Proceedings of the Social Statistics Section*, 695-699.

Seasonal Adjustment of Labour Force Series during Recession and Non-Recession Periods

ESTELA BEE DAGUM and MARIETTA MORRY¹

ABSTRACT

This paper analyzes the revisions of eight seasonally adjusted labour force series during recession and non-recession periods. The four seasonal adjustment methods applied are X-11 and X-11-ARIMA using either concurrent or forecast seasonal factors. The series are seasonally adjusted with these four methodologies according to both a multiplicative and an additive decomposition model. The results indicate that the X-11-ARIMA concurrent adjustment yields the smallest revisions both during recession and non-recession periods regardless of the decomposition model used.

KEY WORDS: Survey; X-11; X-11-ARIMA; Concurrent adjustment; Recession/non-recession.

1. INTRODUCTION

Seasonality in some of the labour force series may be subject to abrupt changes due to dramatic variations in their composition during the various stages of the business cycle. An important example is total unemployment. In relatively prosperous years, it consists mainly of persons shifting jobs, new entrants to the labour market, workers from the primary sector (agriculture, forestry, fishing, trapping, etc.) and construction (in the winter), and students seeking jobs (in the summer). On the other hand, during recessions, the number of unemployed increases quickly and the newly unemployed are mainly regular workers from heavy industries and related activities characterized by seasonal variations of smaller amplitudes and seasonal patterns different from those in 'normal' years. This kind of shift was observed in Canada in 1981-1982, where the total unadjusted unemployment rose from 790,000 in August 1981 to 1,494,000 in December 1982; the newly unemployed coming mainly from the manufacturing and service industries.

The rapid changes in the size and composition of total unemployment during the depressed phase of the business-cycle raises the question as to whether the procedure followed to estimate seasonal factors based on data for years of low, mainly frictional and 'outdoor' unemployment, is applicable to data for years of high unemployment with a large number of the jobless added from the secondary and tertiary sectors.

Empirical research at Statistics Canada in 1974 led to current seasonal adjustment of labour force series by the X-11-ARIMA method using concurrent seasonal factors. This method of adjustment will be referred to as the 'official' procedure in the sections to follow. The U.S. Bureau of Labor Statistics officially adopted the X-11-ARIMA method in 1980 using six-month-ahead projected seasonal factors. This agency also releases monthly the unemployment rate calculated with X-11-ARIMA and concurrent seasonal factors. Concurrent seasonal factors are obtained by seasonally adjusting, each month, all the data available up to and including that month whereas projected seasonal factors are generated from data that ended usually one year before (in the case of the Bureau of Labor Statistics, six-months before).

In Section 2, the mean absolute error (MAE) of concurrent and year-ahead projected seasonal factors is given for eight Canadian labour force series obtained from X-11-ARIMA

¹ Estela Bee Dagum and Marietta Morry, Time Series Research and Analysis Division, Statistics Canada, 13th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

and X-11 using the multiplicative seasonal adjustment option. Year-ahead instead of six-months-ahead projected factors are analyzed because they are applied by several government statistical agencies. Furthermore, the MAE's of six-months-ahead factors fall between those of concurrent and year-ahead projected factors.

The main purpose of this study is to assess whether the use of X-11-ARIMA with concurrent seasonal factors still produces the smallest revisions during recession years when compared to three feasible alternative procedures.

In Section 3, the mean absolute revisions of the additive current seasonal adjustment are calculated for the four alternative procedures and MAE's of the additive are compared to the multiplicative options.

Finally, the conclusions of this study are presented in Section 4.

2. REVISIONS OF CURRENT SEASONALLY ADJUSTED LABOUR FORCE SERIES DURING RECESSION AND NON-RECESSION PERIODS

The majority of the seasonal adjustment methods applied by government statistical agencies are based on linear smoothing filters, usually known as moving averages. It is inherent to these methods that the estimates from the observations of the most recent years are less accurate than those corresponding to central data because of the asymmetry of the end point filters. Among these methods, the Method II-X-11 variant developed by Shiskin, Young, and Musgrave (1967) and X-11-ARIMA developed by Dagum (1980) are the most widely applied. The X-11-ARIMA method is a modified version of the X-11 variant that basically consists of two steps. First, the original series are extended with extrapolation values from ARIMA (autoregressive integrated moving averages) models of the type developed by Box and Jenkins (1970), and then the extended series are seasonally adjusted with a set of moving averages that result from the combination of the X-11 seasonal filters with the extrapolation ARIMA filters. Therefore, the seasonal adjustment filters of X-11-ARIMA and X-11 differ for the data of the most recent year. For both procedures the same symmetric filter is applied to central observations. If the ARIMA option is not used, then the X-11-ARIMA reduces to the X-11 method.

As more data become available, the seasonally adjusted estimate pertaining to a time point keeps getting revised until the data point in question is three years away from the end of the series and the symmetric filters apply, at which point the estimate becomes virtually fixed and is referred to as the final seasonally adjusted estimate. The difference between the very first and the final seasonally adjusted estimate is called the total revision. The revisions of current seasonally adjusted values by the X-11-ARIMA and X-11 methods are due to: (1) Differences in the smoothing linear filters applied to the same observations as more data become available; and, (2) the innovations that enter into the series with new observations. One would like to see the revisions of the first kind reduced to a minimum or completely eliminated.

Theoretical studies by one of the authors (Dagum 1982a and 1982b) have shown that the revisions of current seasonally adjusted values due to filter changes can be reduced substantially if: (1) the original series is extended with ARIMA extrapolated values i.e., the X-11-ARIMA is applied; and (2) concurrent seasonal factors are used instead of year-ahead seasonal factors. The conclusion drawn from these two theoretical studies conforms to the results given in several empirical and theoretical works (see e.g. Dagum 1978, Dagum and Morry 1982, Kuiper 1978, 1981; Pierce 1980; Kenny and Durbin 1982; McKenzie 1982; Wallis 1982; Pierce and McKenzie 1985; Otto 1985).

Next, we examine the performance of X-11-ARIMA with concurrent seasonal factors compared to three other feasible alternatives for recession and non-recession periods. The better seasonal adjustment procedure will be the one that yields smaller revisions.

2.1 Comparisons of Four Alternative Procedures for Current Seasonal Adjustment of Labour Force Series

There are four seasonal adjustment procedures commonly applied to obtain current seasonally adjusted values, namely:

- (1) X-11-ARIMA with concurrent seasonal factors;
- (2) X-11 with concurrent seasonal factors;
- (3) X-11-ARIMA with year-ahead projected seasonal factors; and
- (4) X-11 with year-ahead projected seasonal factors.

The revision measure used here for the evaluation of the four alternative procedures is the mean absolute error (MAE) of the seasonal factors for current seasonal adjustment defined by:

$$MAE(N) = \sum_{t=1}^N |\hat{S}_t^c - \hat{S}_t^F| / N \quad (1)$$

In this expression, N is the number of datapoints included in the mean, denotes the current seasonal factor value which can be either a concurrent or a year-ahead projected seasonal factor from X-11 or X-11-ARIMA. denotes the 'final' seasonal factor in the sense that it will not change significantly when the series is augmented with new data. For X-11 and X-11-ARIMA, a current seasonal factor becomes final when at least three years of data are added to the series (Young 1968; Wallis 1974). This study analyzes the revisions in the seasonal factors (or implicit seasonal factors in the additive case) rather than in the seasonally adjusted estimates for several reasons. First, using seasonal factors provides a feel for the size of revisions relative to the level of the series (it is in the form of a percentage); second, it standardizes the revision size within series subject to substantial jumps in level (such as the unemployment series); third, it allows for cross-series comparisons.

This study analyzes the revisions in the seasonal factors (or implicit seasonal factors in the additive case) rather than in the seasonally adjusted estimates for several reasons. First, using seasonal factors provides a feel for the size of revisions relative to the level of the series (it is in the form of a percentage); second, it standardizes the revision size within series subject to substantial jumps in level (such as the unemployment series); third, it allows for cross-series comparisons.

Unlike in a previous paper by the authors (Dagum and Morry 1982), the revisions in the month-to-month movement of the seasonally adjusted data were not included in the analysis since these revisions are not of primary interest when dealing with labour force data (for example, Statistics Canada does not publish yearly revisions of the growth-rate for these series). Consequently, this paper focuses on the revisions in the level rather than on revisions in the change in level.

The eight Canadian series of employment and unemployment analyzed here start in January 1966 and end in October 1982. To use the ARIMA extrapolation option of X-11-ARIMA a period of at least five years is necessary to produce a seasonally adjusted series. Consequently, the first year for which total revision measures can be calculated is 1971. Taking into account the need for at least three and a half more years for a current estimate to become final, the last full year for which MAE can be obtained is 1977. Within this seven-year span of revisions, we distinguished two years of recession and five years of non-recession. The recession period includes data from August 1974 until July 1975 and June 1976 until May 1977. These two years were considered recessionary because they showed high increases (greater than 25%) in the annual levels of total unemployment due mainly to large inflows of job losers.

Another important aspect taken into consideration is the kind of decomposition model used for the seasonal adjustment of each series. The X-11 and the X-11-ARIMA methods provide both additive and multiplicative decomposition models. There are no theoretical reasons for one model to be preferable to the other. They are based on different assumptions concerning the generating mechanism of the seasonal component.

In an additive model, the components of a time series (trend-cycle, seasonal variations and irregular fluctuations) are assumed to be independent and, therefore, the seasonal effect is not affected by the level of the economic activity conditioned by the stages of the business cycle.

On the other hand, in a multiplicative model, the seasonal effect is proportional to the trend-cycle. If the seasonal factors are constant, it means the higher the level of the seasonally adjusted series, the higher the seasonal effect.

The selection of the decomposition model is not crucial for the estimation of 'final' seasonally adjusted values since for most cases the corresponding figures are similar. The problem of model selection, however, becomes very important when approached from the viewpoint of the estimation of the seasonal component of the end years of a series, particularly, of series with a rapidly growing trend-cycle. The asymmetric filters used for the end points estimation, particularly those of the X-11 method, introduce large systematic errors if the seasonal estimates change fast (Dagum 1978). In fact, if the underlying decomposition model is that of a rather stable multiplicative seasonality, an additive seasonal adjustment will produce seasonal estimates that appear to vary with the trend-cycle. Reciprocally, if stable additive seasonality is the norm, a multiplicative adjustment will produce seasonal factors that look unstable or fast moving.

From the viewpoint of seasonal adjustment, it is then preferable to choose the decomposition model that yields the most stable seasonal estimates. The tests developed by Morry (1975) and Higginson (1977) have been applied to the eight series to determine the preferred decomposition models.

The results of these tests indicated that only two series, unemployment of adult and young women, follow an additive model; the remaining series are of the multiplicative type.

In this study, however, the mean absolute revisions have been analyzed under both assumptions, that is, the components of each series are either multiplicatively or additively related. We are using additive and multiplicative decomposition models for data spanning both recessionary and non-recessionary periods in order to determine which of these two decomposition models is more sensitive to sudden changes of level from the viewpoint of revision.

The calculations shown in the following tables were obtained from multiplicative seasonal adjustment. The results from additive adjustment are discussed in Section 3.

Table 1 shows the mean absolute error (MAE) of the seasonal factors of X-11-ARIMA and X-11 applied for current seasonal adjustment during recession years. It is apparent that X-11-ARIMA with concurrent seasonal factors yields the smallest revisions. This result is consistent with the theoretical findings discussed above which determined that the use of the ARIMA extrapolation option with concurrent seasonal factors significantly reduces filter revisions.

For six out of the eight series analyzed, X-11 with concurrent seasonal factors ranks second. For the other two series (unemployed and employed adult men) X-11/concurrent shows the same MAE results as does X-11-ARIMA with year-ahead projected seasonal factors. Finally, the least accurate estimates are obtained from X-11 with year-ahead projected seasonal factors.

Table 2 shows the relative size of the revisions from each alternative procedure with respect to X-11-ARIMA with concurrent seasonal factors during recession years. All the values are greater than 1.0 indicating that none of the alternative options gives revisions smaller than X-11-ARIMA/concurrent.

Table 1
Mean Absolute Errors (MAE(N)) of Seasonal Factors of X-11-ARIMA
and X-11 during Recession Years^a ($N = 24$)

Series	Concurrent Seasonal Factors		Year-ahead Projected Seasonal Factors	
	X-11-ARIMA (1)	X-11 (2)	X-11-ARIMA (3)	X-11 (4)
Unemployment				
Men 25 +	1.95	2.75	2.74	3.35
Women 25 +	1.94	2.94	3.43	4.70
Men 15-24	2.16	3.02	3.49	4.33
Women 15-24	1.25	1.73	2.48	3.44
Employment				
Men 25 +	0.08	0.12	0.12	0.16
Women 25 +	0.23	0.29	0.33	0.42
Men 15-24	0.41	0.53	0.66	0.76
Women 15-24	0.50	0.70	0.81	0.97

^a August 1974 - July 1975 and June 1976 - May 1977.

Table 2
Comparison of MAE(N)'s from Three Alternative Procedures Versus
X-11-ARIMA/Concurrent for Multiplicative Seasonal Adjustment of Employment
and Unemployment Series in Recession Years ($N = 24$)

Series	X-11 Concurrent vs. X-11-ARIMA Concurrent (1) ^a	X-11-ARIMA Projected Factors vs. X-11-ARIMA Concurrent (2) ^b	X-11 Projected Factors vs. X-11-ARIMA Concurrent (3) ^c
Unemployment			
Men 25 +	1.41	1.40	1.72
Women 25 +	1.52	1.77	2.41
Men 15-24	1.40	1.61	2.00
Women 15-24	1.38	1.98	2.75
Employment			
Men 25 +	1.50	1.50	1.50
Women 25 +	1.26	1.43	1.83
Men 15-24	1.29	1.61	1.85
Women 15-24	1.40	1.62	1.94

^a (1) equals column (2) ÷ column (1) of Table 1.

^b (2) equals column (3) ÷ column (1) of Table 1.

^c (3) equals column (4) ÷ column (1) of Table 1.

The non-recession period covers from January 1971 to December 1977 excluding the recession years. Table 3 shows the MAE of the current seasonally adjusted series for the four procedures during these years. Similarly to Table 1, X-11-ARIMA with concurrent seasonal factors yields the smallest revisions for all the series due to minimal filter revisions as pointed out before. For seven out of the eight series X-11/concurrent ranks second with values relatively close to those shown for X-11-ARIMA with year-ahead projected factors. Finally, the most unreliable procedure in terms of the magnitude of the revision is X-11 with year-ahead seasonal factors.

The relative size of the revisions of the three alternative procedures with respect to the X-11-ARIMA/concurrent procedure during non-recession years are shown in Table 4. The figures in column (1) with the exception of one entry, however, are smaller than those shown in column (1) of Table 2 which would indicate that during recession years the percentage gains achieved by using ARIMA extrapolation are even higher than during non-recession years.

Finally, Table 5 compares the size of the revisions during recession versus non-recession years for the two best procedures. The results show that X-11-ARIMA/concurrent which is Statistics Canada official procedure gives smaller MAE values compared to those of the second best alternative, X-11/concurrent. Most of the ratios in the first column are very close to 1.0, indicating that the revisions in times of recession are similar in size to those in non-recession years when using the ARIMA extrapolation option. If X-11 with concurrent seasonal factors is applied, the size of revision is substantially higher in most series during recession than in 'normal' times. This is due to the fact that the rapid change in the level of the series, introduced by the new observations of the recession years, is not estimated as well by the end filters. In fact, gradual movements and some of the level increase are passed to the seasonal component.

Table 3
Mean Absolute Errors (MAE(N)) of Seasonal Factors of X-11-ARIMA
and X-11 during Recession Years^a ($N = 60$)

Series	Concurrent Seasonal Factors		Year-ahead Projected Seasonal Factors	
	X-11-ARIMA (1)	X-11 (2)	X-11-ARIMA (3)	X-11 (4)
Unemployment				
Men 25 +	1.37	1.73	2.22	2.73
Women 25 +	1.84	2.41	2.92	3.55
Men 15-24	1.97	2.66	3.17	3.96
Women 15-24	1.93	2.87	2.59	3.18
Employment				
Men 25 +	0.08	0.10	0.12	0.13
Women 25 +	0.23	0.27	0.33	0.34
Men 15-24	0.39	0.46	0.58	0.69
Women 15-24	0.43	0.49	0.68	0.80

^a From January 1971 until December 1977 excluding recession periods defined in Table 1 footnote (a)

Table 4

Comparison of MAE(N)'s from Three Alternative Procedures Versus X-11-ARIMA/Concurrent for Multiplicative Seasonal Adjustment of Employment and Unemployment Series in Recession Years (*N* = 60)

Series	X-11 Concurrent vs. X-11-ARIMA Concurrent (1) ^a	X-11-ARIMA Projected Factors vs. X-11-ARIMA Concurrent (2) ^b	X-11 Projected Factors vs. X-11-ARIMA Concurrent (3) ^c
Unemployment			
Men 25 +	1.26	1.62	1.99
Women 25 +	1.31	1.59	1.93
Men 15-24	1.35	1.61	2.01
Women 15-24	1.49	1.34	1.65
Employment			
Men 25 +	1.25	1.50	1.62
Women 25 +	1.17	1.43	1.48
Men 15-24	1.18	1.49	1.77
Women 15-24	1.14	1.58	1.86

^a (1) equals column (2) ÷ column (1) of Table 3.
^b (2) equals column (3) ÷ column (1) of Table 3.
^c (3) equals column (4) ÷ column (1) of Table 3.

Table 5

Comparison of MAE(N)'s of Concurrent Seasonal Factors of X-11-ARIMA and X-11 for Recession Versus Non-Recession Years Using the Multiplicative Option

Series	X-11-ARIMA Concurrent Recession Years (<i>N</i> = 24) vs. Non-Recession Years (<i>N</i> = 60) (1) ^a	X-11 Concurrent Recession Years (<i>N</i> = 24) vs. Non-Recession Years (<i>N</i> = 60) (2) ^b
Unemployment		
Men 25 +	1.42	1.59
Women 25 +	1.05	1.22
Men 15-24	1.09	1.35
Women 15-24	0.67	0.60
Employment		
Men 25 +	1.00	1.20
Women 25 +	1.00	1.07
Men 15-24	1.05	1.27
Women 15-24	1.16	1.54

^a (1) equal to column (1) of Table 1 ÷ column (1) of Table 3.
^b (2) equal to column (2) of Table 1 ÷ column (2) of Table 3.

The only exception is the series unemployed women 15 to 24 where revisions with both methods are smaller during economic hardship. This can be explained by the special behaviour of this series during the period analyzed, which is characterized by large annual increases of about 15% for 1966-73 and 8.5% for 1973-80 and an additive seasonal component, independent of the business-cycle (i.e., the change in level reflected more the changing behaviour of young women than the effect of the business-cycle).

Another special case is the series unemployed men 25 years and over. Here recession years were characterized by much larger revisions than non-recession periods even with ARIMA extrapolations as indicated by a ratio of 1.42. This large discrepancy between the two periods is a result of the drastic composition changes in seasonality that this series undergoes during times of recession as discussed before. Without ARIMA extrapolation, the revision sizes deviate even more (the ratio is 1.59), since apart from the changes in composition the unreliable seasonal estimates produced during recession introduce added discrepancies.

3. COMPARISON OF ADDITIVE VERSUS MULTIPLICATIVE CURRENT SEASONAL ADJUSTMENT DURING RECESSION AND NON-RECESSION PERIODS

It is often argued that during recession periods the use of an additive instead of a multiplicative decomposition model is to be preferred from the viewpoint of the minimization of revisions. The main reasons given for this are: (1) in an additive model, the time series components are assumed to be independent and, therefore, the seasonal effect is not affected by the level of the trend-cycle contrary to what occurs with a multiplicative model; and (2) the inflexibility of the end-point filters to estimate adequately fast-moving seasonality.

The eight labour force series analyzed in the previous section was additively seasonally adjusted in order to assess this new alternative. The results obtained confirm the ranking given by the multiplicative option. Namely, X-11-ARIMA/concurrent yields the smallest revisions followed by X-11/concurrent and X-11-ARIMA/year-ahead projected, in that order. The least accurate estimates are obtained with X-11/year-ahead projected. It is important to note that *factors* of additive seasonal adjustment mean *implicit* factors in the sense that they result from the quotient between the original series and the seasonally adjusted series.

Tables 6 and 7 show the relative size of the revisions by each alternative procedure with respect to X-11-ARIMA/concurrent, for the recession and non-recession periods, respectively. All the values are greater than one indicating that none of the alternative procedures gives smaller revisions than X-11-ARIMA/concurrent. Since the latter ranks first for both additive and multiplicative seasonal adjustment options, we compare for each series which of the two decomposition models gives the smallest revisions.

In Table 8 the data show that for the two series that affect the unemployment rate the most, i.e., the unemployment and employment of adult men, the multiplicative option is to be preferred during recession as well as non-recession years. For the most part, these data confirm the decomposition models chosen by Statistics Canada according to the model tests (Morry 1975; Higginson 1977). The only apparent exception is the series Employed Men 15-24 which would do better with an additive model. However, given the fact that the size of the revisions is already very small, this improvement is of no consequence. The MAE's from the multiplicative adjustment are 0.41 (recession period) and 0.39 (non-recession period) and are reduced by the additive options to 0.33 and 0.31 respectively.

Finally, we observe that the unemployment of adult women would have smaller revisions with a multiplicative instead of an additive seasonal adjustment during recession years.

Table 6

Comparison of MAE(N)'s from Three Alternative Procedures Versus
X-11-ARIMA/Concurrent for Additive Seasonal Adjustment of Employment
and Unemployment Series in Recession Years ($N = 24$)

Series	X-11 Concurrent	X-11-ARIMA Projected Implicit Factors	X-11 Projected Implicit Factors
	X-11-ARIMA Concurrent	X-11-ARIMA Concurrent	X-11-ARIMA Concurrent
Unemployment			
Men 25 +	1.18	1.29	1.38
Women 25 +	1.16	1.49	1.75
Men 15-24	1.21	1.48	1.70
Women 15-24	1.33	1.74	1.84
Employment			
Men 25 +	1.44	1.69	2.08
Women 25 +	1.26	1.33	1.65
Men 15-24	1.02	1.05	1.34
Women 15-24	1.50	1.50	2.05

Table 7

Comparison of MAE(N)'s from Three Alternative Procedures Versus
X-11-ARIMA/Concurrent for Additive Seasonal Adjustment of Employment
and Unemployment Series in Recession Years ($N = 60$)

Series	X-11 Concurrent	X-11-ARIMA Projected Implicit Factors	X-11 Projected Implicit Factors
	X-11-ARIMA Concurrent	X-11-ARIMA Concurrent	X-11-ARIMA Concurrent
Unemployment			
Men 25 +	1.31	1.65	1.88
Women 25 +	1.20	1.59	1.71
Men 15-24	1.22	1.57	1.89
Women 15-24	1.05	1.20	1.26
Employment			
Men 25 +	1.16	1.24	1.54
Women 25 +	1.10	1.27	1.30
Men 15-24	1.22	1.31	1.55
Women 15-24	1.41	1.68	2.16

Table 8

Comparison of MAE(N)'s of Seasonal Factors from Additive Versus Multiplicative X-11-ARIMA (Concurrent) Seasonal Adjustment during Recession and Non-Recession Periods

Series	(N = 24) Recession Period	(N = 60) Non-recession Period
	Additive X-11-ARIMA Concurrent	Additive X-11-ARIMA Concurrent
	Multiplicative X-11-ARIMA Concurrent	Multiplicative X-11-ARIMA Concurrent
Unemployment		
Men 25 +	1.25	1.15
Women 25 +	1.14	0.88
Men 15-24	1.23	1.05
Women 15-24	0.93	0.85
Employment		
Men 25 +	1.25	1.25
Women 25 +	1.00	1.00
Men 15-24	0.80	0.80
Women 15-24	1.14	1.17

4. CONCLUSIONS

The results of Sections 2 and 3 can be summarized as follows:

- (1) The X-11-ARIMA method with concurrent seasonal factors gives the smallest revisions for each series, whether an additive or a multiplicative seasonal adjustment is made, during both recession and non-recession years.
- (2) The comparisons of the magnitude of the revision from additive versus multiplicative seasonal adjustment with X-11-ARIMA/concurrent indicate clearly that the two series that affect the unemployment rate most, unemployment and employment of adult men, are of the multiplicative type during times of recession as well as non-recession.
- (3) During recession years, the use of X-11-ARIMA with year-ahead factors and of X-11/concurrent yields equal MAE's for employment and unemployment adult men. For the six remaining series, however, X-11/concurrent is the second best alternative.
- (4) The least accurate current seasonal adjustment estimates for all series in all the situations discussed are obtained with X-11 with year-ahead projected seasonal factors.
- (5) The comparisons of the revisions during recession versus non-recession periods from X-11-ARIMA/concurrent show that they are of relatively similar magnitude with the important exception being Unemployed Men 25 years and over, where revisions are much higher in recession years. This concurs with the fact that this series undergoes abrupt seasonal changes because of drastic variations in its composition. The larger revisions are mainly due to these new innovations.

On the other hand, the use of concurrent seasonal factors with X-11 shows, for most series, large discrepancies in the size of the revisions of these two periods. This is an indication that revisions result mainly from the inadequacy of the end filters to estimate well the rapidly changing levels of recession periods.

For only one series, Unemployed Women 15-24 years, the two best procedures yield revisions substantially larger in non-recessions compared to recessions. This can be explained by the special behaviour of this series during the analyzed period which is characterized by large annual increases of about 15% for 1966-73 and 8.5% for 1973-80, obscuring the effect of the business-cycle; and, a seasonal component independent of the business-cycle.

Given the above observations, we can feel confident that the official seasonal adjustment procedure at Statistics Canada will give best estimates among the alternatives considered during recession.

ACKNOWLEDGEMENT

The authors are thankful to two anonymous referees whose helpful suggestions contributed to the improvement of this paper.

REFERENCES

- BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.
- DAGUM, E.B. (1978). *Comparison and Assessment of Seasonal Adjustment Methods for Labor Force Series*. Washington, D.C.: U.S. Government Printing Office.
- DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method*. Catalogue 12-564E, Ottawa, Canada: Statistics Canada.
- DAGUM, E.B. (1982a). Revision of time varying seasonal filters. *Journal of Forecasting*, 1, 173-187.
- DAGUM, E.B. (1982b). The effects of asymmetric filters on seasonal factor revisions. *Journal of the American Statistical Association*, 77, 732-738.
- DAGUM, E.B., and MORRY, M. (1982). The estimation of seasonal variations in consumer price indexes. *Proceedings of the Conference on "The Measurement of Prices"*, Catalogue 22-24, Ottawa, Canada: Statistics Canada.
- HIGGINSON, J. (1977). Users manual for the decomposition model test. Research Paper No. 77-01-001, Seasonal Adjustment and Time Series Staff, Statistics Canada.
- KENNY, P., and DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic time series. *Journal of the Royal Statistical Society, Series A*, 145, 1-41.
- KUIPER, J. (1978). A survey and comparative analysis of various methods of seasonal adjustment. *Seasonal Analysis of Economic Time Series* (Ed. Arnold Zellner), Washington, D.C.: U.S. Government Printing Office, 59-76.
- KUIPER, J. (1981). The treatment of extreme values in the X-11-ARIMA program. *Time Series Analysis and Forecasting*, (Eds. Anderson, O., and Perryman, M.R.), Amsterdam: North-Holland Publishing Co., 257-266.
- McKENZIE, S. (1982). An evaluation of concurrent adjustment on Census Bureau time series. *Proceedings of the Business and Economics Section of the American Statistical Association*.
- MORRY, M. (1975). A test for model selection. Research Paper. No. 75-12-016, Seasonal Adjustment and Time Series Staff, Statistics Canada.

- OTTO, M. (1985). Effects of forecasts on the revisions of seasonally adjusted values using the X-11 seasonal adjustment procedure. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association* (forthcoming).
- PIERCE, D. (1980). Data revision with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.
- PIERCE, D., and McKENZIE, S. (1985). On concurrent seasonal adjustment. Technical Paper, U.S. Bureau of the Census.
- SHISKIN, J., YOUNG, A.H., and MUSGRAVE, J.C. (1967). The X-11 variant of census method II seasonal adjustment program. Technical Paper No. 15, U.S. Bureau of Census.
- WALLIS, K.F. (1974). Seasonal adjustment and relations between variables. *Journal of the American Statistical Association*, 69, 18-31.
- WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: Linear filters for the X-11 method. *Journal of the Royal Statistical Society, Series A*, 145, 74-85.
- YOUNG, A.H. (1968). Linear approximations to census and BLS seasonal adjustment methods. *Journal of the American Statistical Association*, 63, 445-457.

Relational Patterns between Total Unemployment and Unemployment Insurance Beneficiaries in Canada

ESTELA BEE DAGUM, GUY HUOT, NAZIRA GAIT,
and NORMAND LANIEL¹

ABSTRACT

This study purports to assess whether there are temporal relationships between Unemployment Insurance Beneficiaries, Total Unemployment, Job Losers and Job Leavers in Canada using univariate and multivariate time series methods. The results indicate that during 1975-82 the Unemployment Insurance Beneficiaries series leads: (1) Total Unemployment by one month and (2) Job Leavers by two months. On the other hand, there are evidence of a feedback relationship between Unemployment Insurance Beneficiaries and Job Losers.

KEY WORDS: Job losers; Job leavers; ARIMA; VARMA; Multivariate time series.

1. INTRODUCTION

Unemployment Insurance (UI) plays a key role in helping the national labour markets adjust to trade and demand-induced changes in production and employment patterns. The main function of UI as part of labour market policy is to provide adequate financial protection during temporary unemployment, to facilitate adjustments. By removing the immediate threat from unemployment, UI relieves job seekers of the need to yield to economic pressures by accepting jobs unsuited to their skills or abilities. It permits a more systematic or wide-ranging job search contributing to the efficient reallocation of human resources. Furthermore, when there are temporary plant layoffs, the objective of UI is met by providing income protection to laid-off workers, so the employer keeps an experienced labour force intact. This saves him/her the cost of recruiting and training new employees after a layoff. It also saves the employee from going through extreme dislocation to prevent financial hardship.

In any situation, UI must have enough flexibility to take into account prevailing economic circumstances which may limit the availability of other jobs and extended jobseekers' unemployment. In the Canadian UI program, this flexibility is provided as longer benefit durations are triggered by rising regional unemployment rates.

The gap between overall unemployment and the UI series tends to narrow in recession and widen in recovery periods. Where business conditions worsen and layoffs occur, job losers become a greater proportion of Total Unemployment. As the most Unemployment Insurance claimants are in fact job losers, this increases the proportion of Unemployment Insurance Beneficiaries related to Total Unemployment.

This study purports to assess whether there is a temporal relationship between the Unemployment Insurance Beneficiaries and Total Unemployment in Canada. The analysis is extended to Job Losers (JLo) and Job Leavers (JLe) who can claim for benefits and are the two major groups of Total Unemployment. The existence of strong relationships among these variables can be useful to explain labour markets behaviour. Furthermore, they may lead to other types of similar relationships useful to estimate unemployment in small areas

¹ E.B. Dagum and G. Huot, Time Series Research and Analysis Division, Statistics Canada. N. Gait, University of Sao Paulo, Brazil, was visiting Statistics Canada when the paper was written, and N. Laniel, previously Time Series Research and Analysis Division, currently with Business Survey Methods Division, Statistics Canada.

where the sample size of the current labour force survey is inadequate. Section 2 introduces the definition of each of the four series discussed and analyzes the main characteristics from their spectra. Section 3 estimates the residual cross-correlation values, for several time lags, of the whitened series to assess whether or not there are pairwise relationships and their direction, if present. The residuals are computed from ARIMA models fitted to each series. Section 4 extends the previous analyses by identifying and estimating two multivariate time series models in order to understand the joint dynamic relationships of: (1) UIB and TU; and (2) UIB, JLo and JLe. Finally, Section 5 gives the main conclusions of this study.

2. THE MAIN CHARACTERISTICS OF THE ANALYZED SERIES

To understand the type of relationship between UIB and TU and its major components, JLo and JLe, we first introduce the definitions and analyze the main characteristics looking at their spectra.

2.1 Total Unemployment (TU)

The Labour Force Survey (LFS) Division of Statistics Canada obtains monthly information through a sample of 56,000 representative households across the country. Although developed since 1952, substantial revisions were introduced to the LFS from 1976.

Estimates of employment, unemployment and non-labour force activity refer to the specific week covered by the survey each month, normally the week containing the 15th day. The sample is designed to represent all persons in the population 15 years of age and over, residing in Canada, with some minor exceptions.

The Labour Force is composed to people who, during the reference week, were employed or unemployed. The employed includes persons who:

- did any work at all;
- had a job but were not at work due to illness or disability, bad weather, labour dispute, vacation, personal or family responsibilities.

The unemployed includes persons who:

- were without work, but actively looked for work in the past four weeks and were available for work;
- had not actively looked for work in the past four weeks but had been on layoff for 26 weeks or less, and were available for work;
- had not actively looked for work in the past four weeks but had a new job to start in four weeks or less, and were available for work.

Total unemployment is composed of the sum of job losers (JLo), job leavers (JLe), new entrants to the labour market, re-entrants after one year or less, re-entrants after more than one year (Statistics Canada 1976). Of these five components, the first two are the most important for our study since they can claim benefits and represent about 70% of TU.

Data on the flows into unemployment are not available prior to 1975. Thus, all the series were observed for the period January 1975 to December 1982, thus including the most recent data available at the time.

Figure 1 shows the original Total Unemployment series which is characterized by a peak in the winter months and a trough in the summer. Figure 2 shows the spectrum of the Total Unemployment series. High power is observed at the frequency 0.05 cycle/month associated with the business-cycle (0.05 corresponds to a 20-months cycle). Similarly, relatively high power is observed at the fundamental seasonal frequency 0.083 cycle/month and neighbouring frequencies, but less at the harmonics of the fundamental seasonal. Finally, the contribution of the irregular fluctuations to the total variance is small, relative to the other two components.

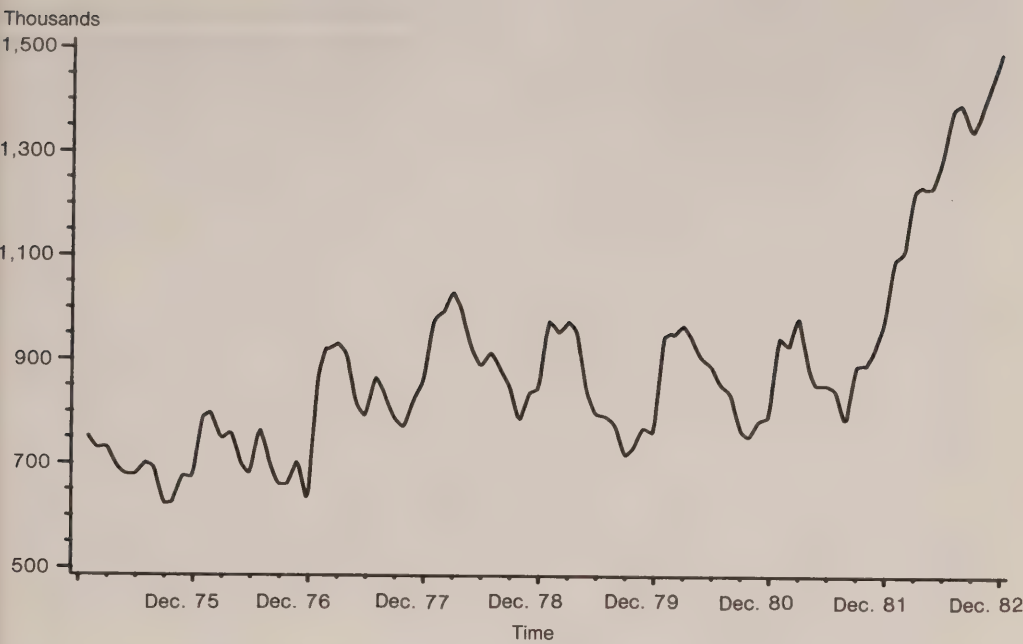


Figure 1. Total Unemployment Series

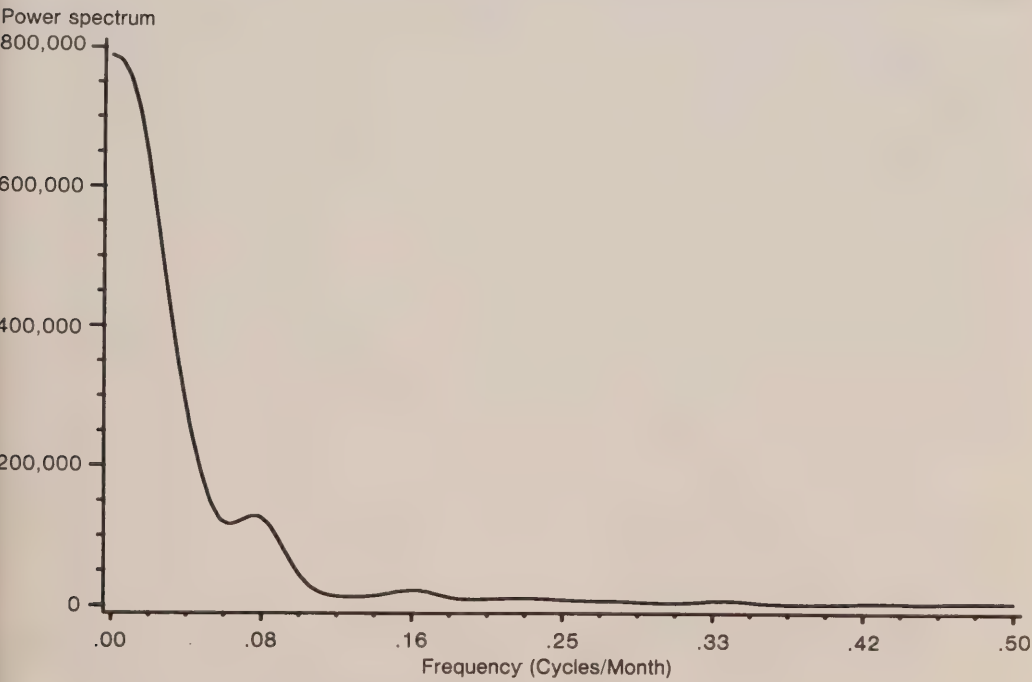


Figure 2. Spectrum of Total Unemployment

Figure 3 shows the original Job Losers series and Figure 4 displays its corresponding spectrum. Similar to TU, high power is shown at the business-cycle frequencies, but now most of the seasonal power is at the fundamental seasonal band and very little is left at the harmonic bands. The contribution of the irregular variations is smaller than that of TU.

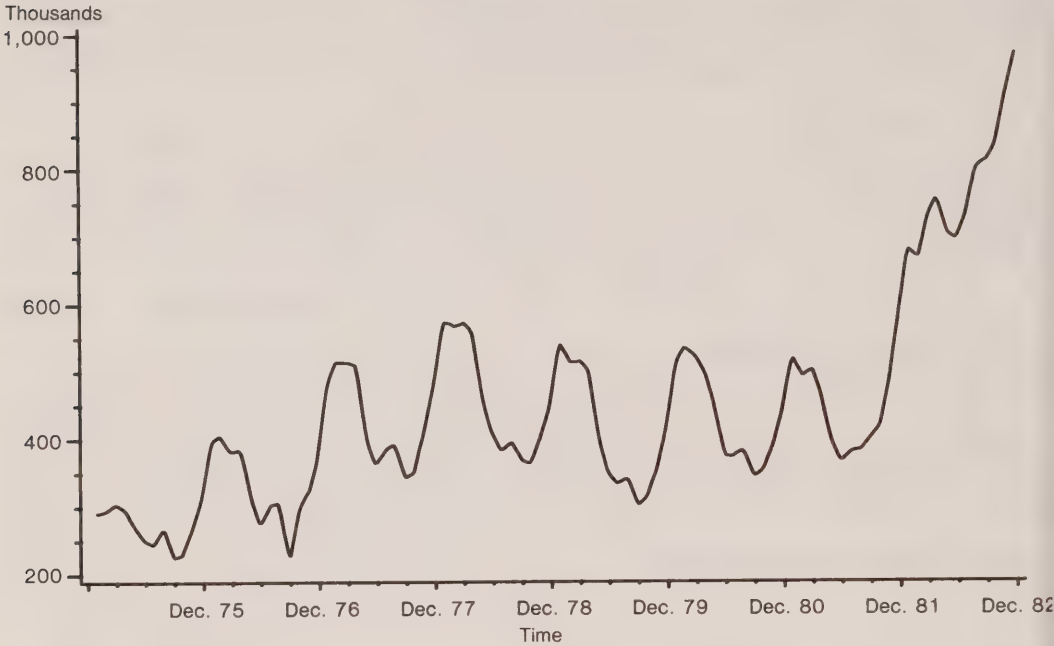


Figure 3. Job Losers Series

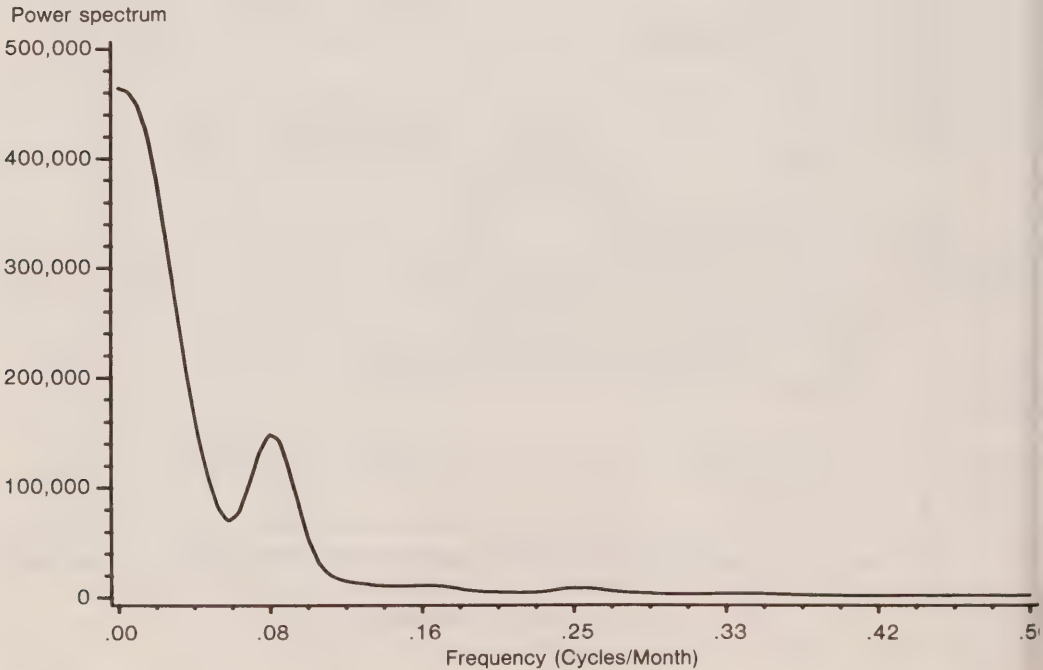


Figure 4. Spectrum of Job Losers

Figure 5 shows the Job Leavers series, characterized by two troughs, one in the winter months and the other during the summer. Its spectrum is given in Figure 6. This series has more cyclical variations than trend as indicated by the high peak at 0.022 cycle/month which corresponds to a 45 months-cycle. Furthermore, the seasonal variations are highly concentrated around the first harmonic band, supporting the fact that this series has two seasonal troughs. Finally, the contribution of the irregular to the total variance is larger than that observed for the two previous series.

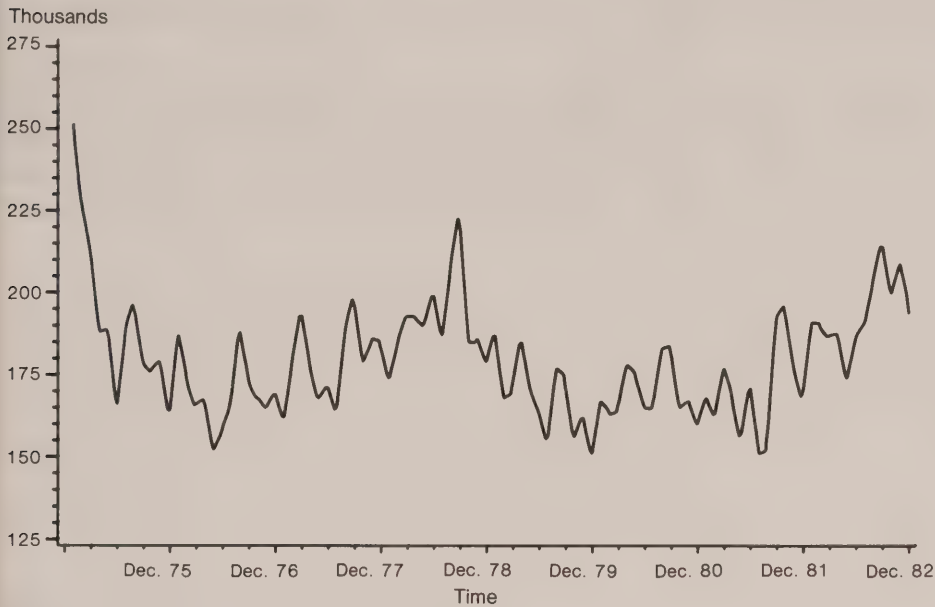


Figure 5. Job Leavers Series

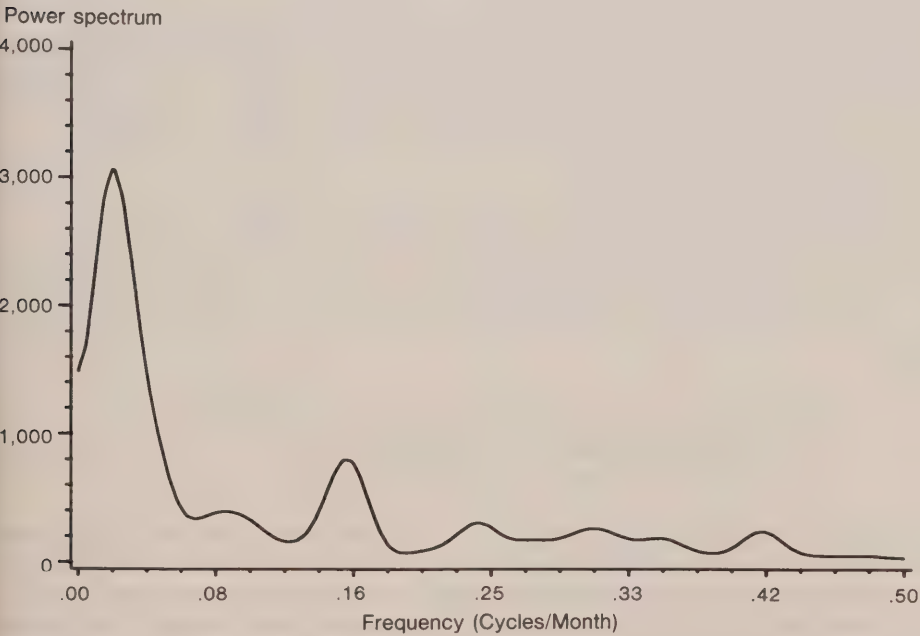


Figure 6. Spectrum of Job Leavers

2.2 The Unemployment Insurance Beneficiaries (UIB)

The monthly data for Unemployment Insurance Beneficiaries cover all persons drawing benefits for a specific week, namely the week of the LFS. This is not a sample since it includes the total population of beneficiaries. The UI covers virtually all paid workers in the labour force and members of Armed Forces. The main exceptions are:

- People 65 years of age and over;
- People working fewer than 15 hours weekly;
- People earning less than 20% of the maximum weekly insurable earnings (in 1982, it was \$70).

In order to qualify for benefits, a claimant must be available for and capable of work, unable to find suitable employment and have the necessary qualifying requirements. Previously eight weeks of work was the minimum required to qualify for benefits but as of December 1977 this number varied between 10 and 14 weeks according to the rate of unemployment prevailing in the region of residence of the claimant. Benefits are paid after a two-week period has been served.

Claimants who qualify for benefits can receive up to 25 percent of their benefits in earnings and continue to receive UI. However, the LFS considers these individuals to be employed. In order to assess the relationship between UI beneficiaries and unemployment, it is thus more accurate to use the series of UI beneficiaries *without* earnings. This subset of UI beneficiaries is a fairly consistent and significant proportion of the total LFS count of the unemployed. We must note, however, that because of differences in definition, the following groups are counted as unemployed in the LFS but are not included in the UI records, namely, entrants and re-entrants; all individuals who have worked but not long enough to qualify for benefits; and those unemployed persons who were previously self-employed. On the other hand, persons insured under the UI program can receive benefits even though, under the LFS definition they would not be classified as unemployed, examples include self-employed fishermen during the off-season, women on maternity leave and employees away from work due to sickness or disability.

The UI beneficiaries (without earnings) series is a sensitive indicator of labour market economic conditions. It is reflective of the insured labour force with recent work experience.

The original Unemployment Insurance Beneficiaries series, as shown in Figure 7, displays large seasonal fluctuations with a peak during the winter months, when bad weather curtails outdoor work in such industries such as fishing, construction and lumber, bringing a sharp rise in claims filed by affected workers.

Figure 8 shows the spectrum of the UIB series. Very high power is shown at the frequency 0.0167 cycle/month, which corresponds to a 60 months-cycle, and at those frequencies associated with the fundamental seasonal band. The contribution of seasonal variations to the total variance of the series is much larger than that observed in TU and its two major components. Finally, there is little irregularity relative to the trend-cycle and the seasonal components.

3. PAIRWISE RELATIONSHIPS BETWEEN UNEMPLOYMENT INSURANCE BENEFICIARIES, TOTAL UNEMPLOYMENT, JOB LOSERS AND JOB LEAVERS

Several early Canadian studies (e.g., Grubel *et al.* 1975; Green and Cousineau 1976; Jump and Rea 1975; and Siedule *et al.* 1976) support the general conclusion that unemployment has tended to shift upward with the increased availability of unemployment insurance in 1971. Lazar (1978) shows that the 1971 changes increased the unemployment duration and induced higher rates of job leaving, especially of young persons and adult women. These studies

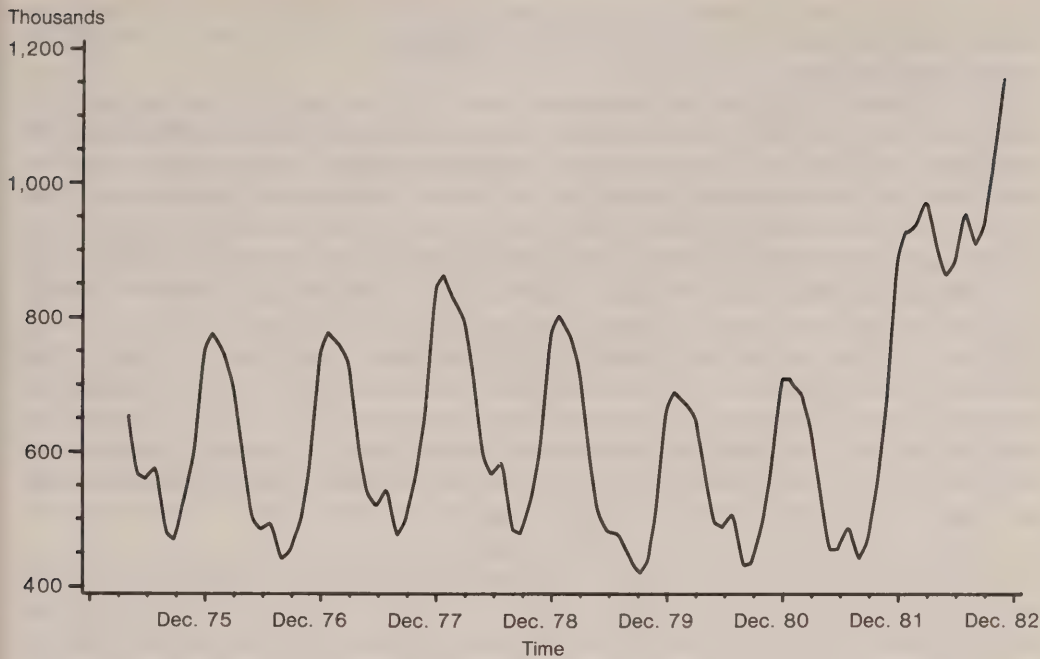


Figure 7. Unemployment Insurance Beneficiaries Series

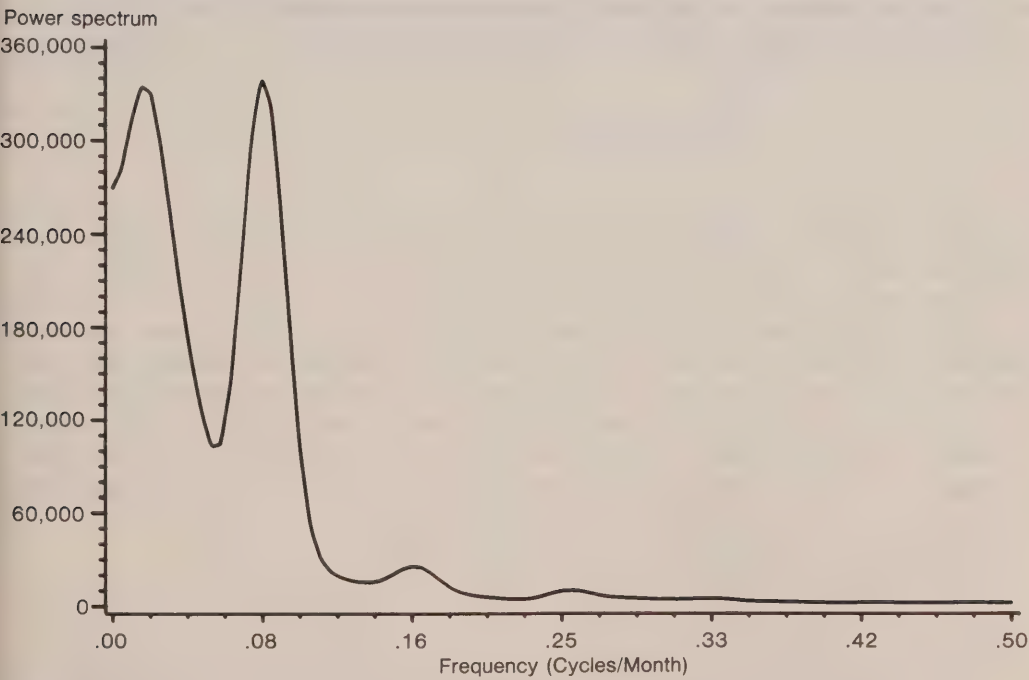


Figure 8. Spectrum of Unemployment Insurance Beneficiaries

were made before the changes of 1975 that aimed at strengthening work incentives. It was expected that the changes introduced after 1975 would reverse the effects of the program on total unemployment.

In this section, we carry out an exploratory analysis by searching for pairwise temporal relationships between Total Unemployment, Unemployment Insurance Beneficiaries, Job Losers and Job Leavers. The existence of these relationships will be useful to build a multivariate time series model to explain the joint dynamic behaviour of the above variables.

The pairwise relationships between TU, UIB, JLo and JLe are calculated using the cross-correlations of the residuals or *innovations* from ARIMA models (Box and Jenkins 1970) that fitted well the data. It has been rightly argued by several authors (e.g., Pierce and Haugh 1977) that the cross-correlations between white noise residuals obtained with different filters are biased to accepting the nul hypothesis of independence when it does not exist. Pierce and Haugh (1977) suggest to use dynamic regression models. This, however, implies that we have to make a judgement on which variable is the *cause* and which is the *effect*. At this stage, we are simply interested in determining whether there is a temporal relationship in each pair of variables analyzed. Table 1 shows the ARIMA models fitted to each series, their parameter values estimated with unconditional least squares, the results of the portmanteau test (Ljung and Box 1978) and the residual variance.

The Q statistics values accept the null hypothesis of randomness of the residuals in each case. However, since this test is applied to a set of autocorrelations of residuls for various lags, it is possible to have significant autocorrelation for some particular time lag k that will not be detected by this test. Therefore, we also tested whether there was autocorrelation of the residuals for each time lag. We used a more accurate approximation for small samples than $1/N$ to test the variance of the autocorrelation, that is, $(N - |k|)N^{-2}$ as given by Haugh (1976).

Having obtained satisfactory results from the above models we calculated the cross-correlation $\hat{r}_{xy}(k)$ between the series analyzed. The S_M^* statistic (Haugh 1976) is applied to test the independence between the series. Under the assumption that the residuals are normally distributed and that $E[\hat{r}_{xy}(k)] = 0$ and $\text{Var}[\hat{r}_{xy}(k)] = (N - |k|)N^{-2}$, the statistic

$$S_M^* = N^2 \sum_{k=-M}^M (N - |k|)^{-1} \hat{r}_{xy}(k)^2$$

follows a X^2 distribution with $2M + 1$ degrees of freedom. In order to determine the direction of the pairwise relationships, we modified the S_M^* statistics which is calculated for positive or negative k only, excluding zero.

Table 2 presents the estimates of the cross-correlation between Unemployment Insurance Beneficiaries (UIB) and Total Unemployment (TU) and its two major subcomponents Job Losers (JLo) and Job Leavers (JLe). We indicate with (a) and (b) those values significant at a 5% and 1% confidence level. In the case of UIB and JLo we calculated S_M^* for positive and negative values of k from ± 1 to ± 6 and from ± 1 to ± 2 to determine whether there is a dominant unidirectional relationship. The results indicated that there is no dominant direction between the two variables but a feedback process.

We can summarize the results from Table 2 as follows:

- (1) There is indication of a unidirectional relationship between UIB and TU such that UIB would lead TU by one month;
- (2) There is a feedback between UIB and JLo with a strong instantaneous relationship. Taking into consideration the time lag between the two variables, the feedback process seems to be initiated by JLo at lag 2.

Table 1
Univariate ARIMA Models

Series	ARIMA Models	Q(24)	$\hat{\sigma}^2_a$
Unemployment Insurance Beneficiaries (UIB)	$(1 - 0.68B)\Delta\Delta^{12} \log_{10} UIB_t = (1 - 0.80B^{12})a_t$	11.55	0.000140
Total Unemployment (TU)	$(1 - 0.25B^3)\Delta\Delta^{12} \log_{10} TU_t = (1 - 0.84B^{12})a_t$	9.13	0.000395
Job Losers (JLo)	$(1 - 0.31B^3)\Delta\Delta^{12} \log_{10} JLo_t = (1 - 0.67B^{12})a_t$	15.78	0.000604
Job Leavers (JLe)	$(1 - 0.37B^3)\Delta\Delta^{12} \log_{10} JLe_t = (1 - 0.40B - 0.25B^2)(1 - 0.87B^{12})a_t$	14.58	0.000627

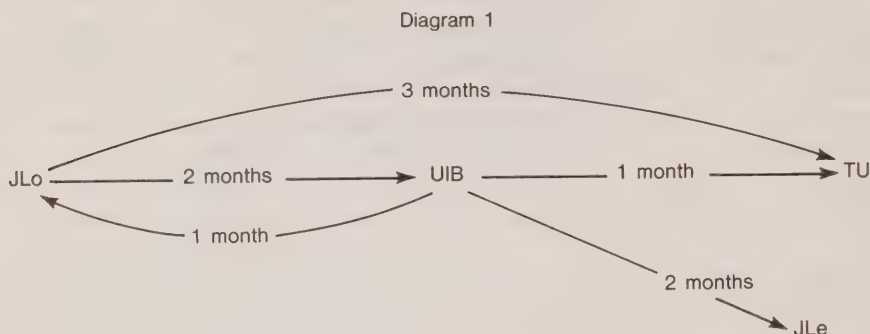
Table 2
Cross-Correlation Between Unemployment Insurance Beneficiaries and Total Unemployment and Its Two Major Components, Job Losers and Job Leavers

LAGS k	$UIB_{(t-k)} - TU_t$ $\hat{r}(k)$	$UIB_{(t-k)} - JLo_t$ $\hat{r}(k)$	$UIB_{(t-k)} - JLe_t$ $\hat{r}(k)$	$JLo_{(t-k)} - TU_t$ $\hat{r}(k)$
-6	-0.07	0.01	0.27 ^a	0.15
-5	0.05	-0.04	-0.09	-0.04
-4	-0.09	0.03	-0.08	-0.01
-3	0.01	0.01	-0.11	-0.06
-2	0.14	0.28 ^a	0.14	-0.01
-1	0.14	0.08	-0.01	0.21
0	0.16	0.32 ^b	0.06	0.39 ^b
1	0.22 ^a	0.29 ^a	0.04	0.14
2	0.12	0.12	0.26 ^a	-0.16
3	-0.07	0.00	0.19	0.42 ^b
4	0.12	-0.05	0.01	-0.05
5	-0.06	0.09	0.11	-0.04
6	0.13	0.00	0.08	0.05

^a 5% significance level.
^b 1% significance level.

- (3) There is a unidirectional relationship between UIB and JLe such that UIB would lead JLe by 2 months. We observe, however, the effect of a delayed feedback at lag 6 which arises from the fact that the JLe series have a strong secondary peak in summer as shown in Figure 6.
- (4) Finally, there is a strong instantaneous and unidirectional relationship between JLo and TU such that JLo would lead TU by 3 months.

The above observations lead to the following Diagram 1 which will be useful for the identification of a more complex multivariate time series model that also takes into account the partial associations among the variables.



4. BUILDING A MULTIVARIATE TIME SERIES MODEL FOR UNEMPLOYMENT INSURANCE BENEFICIARIES, TOTAL UNEMPLOYMENT, JOB LOSERS AND JOB LEAVERS

In the previous section we concluded that there are pairwise relationships among the four variables in the sense defined by Granger (1969) and Pierce and Haugh (1977). Taking into consideration those preliminary relationships, we here identify and estimate two multivariate time series models following the methodology developed by Tiao and Box (1981) and Tiao and Tsay (1983). These models will explain the joint dynamic behaviour of the variables involved.

A vector ARMA model for seasonal series takes the form

$$\phi(B)\Phi(B^s)\underline{Z}_t = \underline{\theta}(B)\Theta(B^s)\underline{a}_t \quad (4.1)$$

where

$$\phi(B) = \underline{I} - \phi_1 B - \dots - \phi_p B^p \quad (4.2)$$

$$\Phi(B^s) = \underline{I} - \Phi_1 B^s - \dots - \Phi_p B^{sp} \quad (4.3)$$

$$\underline{\theta}(B) = \underline{I} - \underline{\theta}_1 B - \dots - \underline{\theta}_q B^q \quad (4.4)$$

$$\Theta(B^s) = \underline{I} - \Theta_1 B^s - \dots - \Theta_Q B^{sQ} \quad (4.5)$$

are matrix polynomials in B (the back shift operator which is defined by $B^m \underline{Z}_t = \underline{Z}_{t-m}$), the ϕ 's, Φ 's, $\underline{\theta}$'s and Θ 's are $k \times k$ matrices, s is the seasonal periodicity and \underline{a}_t is a sequence of random shock vectors $\text{IID } N(\underline{0}, \underline{\Sigma})$ and \underline{Z}_t is a vector of stationary time series.

In order to avoid a problem of multicollinearity between TU and JLo, two VARMA models were specified, a VARMA (1,2)(0,1)₁₂ that relates Unemployment Insurance Beneficiaries with total Unemployment, and a VARMA (2,6)(0,1)₁₂ that relates UIB with Job Losers and Job Leavers. These models were identified and estimated using the exact maximum likelihood method in the Scientific Computing Associates program (Liu and Hudak 1983). The models are fitted respectively to the original data transformed as follows:

$$\begin{pmatrix} uib_t \\ tu_t \end{pmatrix} = (1 - B)(1 - B^{12}) \log_{10} \begin{pmatrix} UIB_t \\ TU_t \end{pmatrix} \tag{4.6}$$

and,

$$\begin{pmatrix} uib_t \\ jlo_t \\ jle_t \end{pmatrix} = (1 - B)(1 - B^{12}) \log_{10} \begin{pmatrix} UIB_t \\ JLo_t \\ JLe_t \end{pmatrix} \tag{4.7}$$

Table 3 shows the parameter values of the VARMA (1,2)(0,1) model and the standard errors of estimates given in parenthesis. (The estimated parameter values and the variance-covariance matrix of the residuals shown in Table 3 cannot be compared with the one of the univariate models (Table 1) because the former result from the fit of the model to the standardized transformed data instead of the non-standardized as it was the case with the univariate models.) Examination of the pattern of the cross-correlations of the residuals in Table 4 suggests that the model is adequate. A plus (minus) sign is used when the estimate is greater (less) than twice its standard error and a dot for a non-significant value based on the above criterion.

Thus, the VARMA model for UIB and TU becomes,

$$uib_t = 0.669uib_{t-1} + \hat{a}_{1t} - 0.794 \hat{a}_{1(t-12)} \tag{4.8}$$

$$\begin{aligned} tu_t = & 0.475uib_{t-1} - 0.347tu_{t-1} + \hat{a}_{2(t)} - 0.308\hat{a}_{2(t-2)} \\ & - 0.705\hat{a}_{2(t-12)} + 0.217\hat{a}_{2(t-14)} \end{aligned} \tag{4.9}$$

Table 3
Estimated Parameters for the Transformed UIB and TU Variables

$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_{12}$
$\begin{bmatrix} 0.669 & - \\ (0.089) & \end{bmatrix}$	$\begin{bmatrix} - & - \\ - & 0.308 \\ & (0.116) \end{bmatrix}$	$\begin{bmatrix} 0.794 & - \\ (0.090) & \\ - & 0.705 \\ & (0.086) \end{bmatrix}$
$\hat{\Sigma}$	$\hat{\theta}_1$	
$\begin{bmatrix} 0.429249 & - \\ 0.131532 & 0.544389 \end{bmatrix}$	$\underline{0}$	

Table 4
Cross-Correlation Matrices of the Residuals in Terms of +, -, and .

LAGS 1 THROUGH 6					
..
..
LAGS 7 THROUGH 12					
..
..
LAGS 13 THROUGH 18					
..
..
LAGS 19 THROUGH 24					
..
..

Equations (4.8) and (4.9) indicate that Unemployment Beneficiaries leads the Total Unemployment series by one month. In fact, when analyzing the relationship between UIB and TU we must keep in mind that an increase in JLo and thus an increase in UIB may lead other members of the family to look for work in order to compensate for the loss of income. These are the new entrants and re-entrants who do not qualify for insurance benefits but contribute to an increase in TU. Furthermore, we should note that it is possible to have an increase in Total Unemployment without an increase in the normal gross flow of labour markets, simply because an increase in UIB occurs during recessionary periods where the availability of jobs is significantly reduced and thus flows into the unemployment state will increase.

The results of this model are in agreement with the preliminary results obtained from the pairwise cross-correlations of the previous section as shown in Diagram 1. The model, however, provides us with a more complete information on the dynamic behaviour of these two phenomena. We observe that the Unemployment Insurance Beneficiaries series is positively related to its previous-month level whereas the Total Unemployment is positively related to the previous-month level of UIB and negatively related to its previous-month level. In both equations, the effect of seasonality is reflected in their moving average part with a high parameter value for the random shock at lag 12.

Table 5 shows the VARMA (2,6)(0,1)₁₂ model applied to the transformed UIB, JLo and JLe variables as given in System (4.7).

Table 6 indicates no recognizable patterns in the estimated cross-correlation matrices of the residuals and, therefore, this model is considered adequate.

The final vector ARMA (2,6)(0,1)₁₂ model for the three variables is,

$$\begin{aligned} uib_t = & 0.617uib_{t-1} + 0.268jlo_{t-2} + \hat{a}_{1(t)} \\ & + 0.221\hat{a}_{3(t-6)} - 0.831\hat{a}_{1(t-12)} - 0.176\hat{a}_{3(t-18)} \end{aligned} \quad (4.10)$$

$$\begin{aligned} jlo_t = & 0.577uib_{t-1} - 0.285jlo_{t-1} + \hat{a}_{2(t)} \\ & + 0.386\hat{a}_{3(t-6)} - 0.525\hat{a}_{2(t-12)} - 0.308\hat{a}_{3(t-18)} \end{aligned} \quad (4.11)$$

$$\begin{aligned} jle_t = & 0.303uib_{t-2} - 0.411jle_{t-1} - 0.403jle_{t-2} \\ & + \hat{a}_{3(t)} - 0.797\hat{a}_{3(t-12)}. \end{aligned} \quad (4.12)$$

Table 5
Estimated Parameters for the Transformed UIB, JLo and JLe Variables

$\hat{\phi}_1$			$\hat{\phi}_2$			$\hat{\theta}_6$		
0.617 (0.086)	-	-	-	0.268 (0.080)	-	-	-	-0.221 (0.087)
0.577 (0.099)	-0.285 (0.096)	-	-	-	-	-	-	-0.386 (0.108)
-	-	-0.411 (0.088)	0.303 (0.083)	-	-0.403 (0.084)	-	-	-

$\hat{\phi}_{12}$			Σ			$\hat{\theta}_t, t = 1,2,\dots,5$		
0.831 (0.094)	-	-	0.339	-	-	<u>0</u>		
-	0.525 (0.096)	-	0.117	0.483	-			
-	-	0.797 (0.077)	0.014	0.153	0.428			

Table 6
Cross-Correlation Matrices Terms of +, -, and ·

LAGS 1 THROUGH 6					
...
...
...
LAGS 7 THROUGH 12					
..+
...
...
LAGS 13 THROUGH 18					
..-	-..	...
...
...
LAGS 19 THROUGH 24					
...
...
...

Equation (4.10) and (4.11) shows the existence of feedback between Job Losers and Unemployment Insurance Beneficiaries similar to the relationship found in section 3. The JLo series leads UIB by two months (equation 4.10) and the one month lagged UIB strongly affects the current value of JLo (equation 4.11). Furthermore, each of the two endogenous variables UIB and JLo are affected by their previous-month levels, positively in the case of UIB and negatively in the case of JLo. The relationship between both series due to seasonality is reflected by the parameter values of a_{t-6} and a_{t-12} . The need for a moving average term at lag 6 arises from the fact that the JLe series have a strong secondary peak in summer as shown in figure 6.

These empirical results are not in contradiction with economic theory. It has been argued, with good reason, that causality cannot be detected only from empirical evidences but must be supported by economic theory (see e.g. Zellner 1979). It is easy to accept that an increase in Job Losers which is associated with an economic recession will lead to an increase in Unemployment Insurance Beneficiaries. In turn, an increase in Unemployment Insurance Beneficiaries will lead to an increase in Job Losers because in reaction to a severe economic recession, most firms make temporary layoffs to be able to have their employees back when economic conditions improve.

Equation (4.12) raises an interesting question when showing that Unemployment Insurance Beneficiaries leads the Job Leavers by two months. It is not so evident why this should be the case.

Plausible explanations can be found in the analysis of the shortrun dynamics of the Canadian labour markets and a thorough investigation would require longitudinal data. We can, however, entertain the hypothesis among others that an increase in JLo and thus an increase in UIB may lead other members of the family to look for work in order to compensate for the loss of income. These persons are the new entrants and re-entrants. During a recession when JLo is increasing it is very difficult for new entrants and re-entrants to find a job. These new entrants and re-entrants are mainly young people and women over 25 who are willing to accept any job, at first, as long as it means extra income for the family. They might work for the length of time necessary for them to qualify for benefits. Then, once they qualify for benefits, they would become JLe in order to be more selective in the kind of job they will accept.

5. CONCLUSIONS

The main purpose of this study has been to assess whether there are temporal relationships between Unemployment Insurance Beneficiaries (UIB) and Total Unemployment (TU), Job Losers (JLo) and Job Leavers (JLe) by building dynamic multivariate time series models.

We have first carried out an exploratory analysis by searching for pairwise temporal relationships between TU, UIB, JLo and JLe in the sense defined by Granger (1969) and Pierce and Haugh (1977). Our results indicated the existence of relationships among the four variables involved.

We have then identified and estimated two multivariate time series models following the methodology developed by Tiao and Box (1981) and Tiao and Tsay (1983). The results of the vector ARMA models agree with the preliminary results obtained from the pairwise cross-correlations of the residuals of the univariate ARIMA models.

The first vector ARMA model shows that the UIB series leads TU by one month. UIB is also positively related to its previous-month level whereas TU is negatively related.

The second vector ARMA model shows that JLo leads UIB by two months with the existence of a one-month feedback from UIB to JLo. Furthermore, UIB is positively affected by its previous-month level while JLo is negatively related. It also shows that UIB leads JLe by two months.

These empirical results based on data for 1975-82 are not in contradiction with economic theory. Furthermore, they conform to those of earlier Canadian studies, based on data prior to 1975, which supported the general conclusions that the increased availability of unemployment insurance induced higher rates of job leaving, especially of young persons and adult women and led to increased levels of unemployment. Hence, it seems that the UIC regulation change in 1977 had little effect, if any, in this regard.

It would have been very interesting to assess the effect of the high recession that started in July 1981 but given the series length, elimination of this recessionary period would have made the series too short for any sound statistical modelling.

REFERENCES

- BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis Forecasting and Control*. San Francisco: Holden Day.
- GRANGER, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral Methods. *Econometrica*, 37, 424-438.
- GREEN, C., and COUSINEAU, J.M. (1976). Unemployment in Canada: The impact of unemployment insurance. *The Economic Council of Canada, Ottawa*.
- GRUBEL, H.G., MAKI, D., and SAX, S. (1975). Real and insurance induced unemployment in Canada. *Canadian Journal of Economics*, VIII, 174-191.
- HAUGH, L.D. (1976). Checking the independence of two covariance stationary time series: A univariate residual cross-correlation approach. *Journal of American Statistical Association*, 71, 378-385.
- JUMP, G.V., and REA, S.A. (1975). The impact of the 1971 unemployment insurance act on work incentives and the aggregate labour market. *Institute for Policy Analysis*, University of Toronto.
- LAZAR, F. (1978). The impact of the 1971 unemployment insurance revisions on unemployment rates: Another look. *Canadian Journal of Economics*, August, 559-570.
- LIU, L.M., and HUDAK, G.B. (1983). *Univariate - Multivariate Time Series and General Statistical Analysis*. Illinois: Scientific Computing Associates.
- LJUNG, G.M., and BOX, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-307.
- PIERCE, D.A., and HAUGH, L.D. (1977). Causality in temporal systems. *Journal of Econometrics*, 5, 265-293.
- SIEDULE, T., SKOULAS, N., and NEWTON, K. (1976). The impact of economy-wide changes on the labour force, an econometric analysis. *The Economic Council of Canada, Ottawa*.
- STATISTICS CANADA (1976). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-716.
- TIAO, G.C., and BOX, G.E.P. (1981). Modelling multiple time series with applications. *Journal of American Statistical Association*, 76, 802-816.
- TIAO, G.C., and TSAY, R.S. (1983). Multiple time series modelling and extended sample cross-correlations. *Journal of Business and Economic Statistics*, 1, 43-56.
- ZELLNER, A. (1979). Causality and econometrics. *Carnegie Rochester Conference Series on Public Policy*, Volume 10, (Eds. Karl Brunner and Allan H. Meltzer), Amsterdam: North-Holland Publishing Company, 9-54.

Basic Principles of Questionnaire Design

LARRY SWAIN¹

ABSTRACT

Thirty basic principles of questionnaire design are presented covering the content, wording, format, and testing of questionnaires. The extent to which the questionnaire is an integral part of the survey is emphasized as is consideration of its relationship with other aspects of survey design.

KEY WORDS: Survey; Questionnaire; Methodology.

1. INTRODUCTION

Most surveys make use of a questionnaire which is to be completed by either a respondent or an official representative of a survey organization (by personal contact or telephone). Since the questionnaire is the means by which the objectives of a survey are transformed into measurable variables, successful achievement of those objectives requires an effective questionnaire. In addition, the questionnaire may help structure, standardize, and control the data collection process so that the required information is obtained in a satisfactory manner. Effective questionnaire design is a combination of basic principles and common sense, adapted to the particular needs of each individual survey.

Although thirty separate principles of questionnaire design are presented, they are not intended to be seen as independent of each other or of the survey environment in which they operate. The extent to which the questionnaire is an integral part of the survey process cannot be sufficiently emphasized. As the questionnaire *cannot* be designed in isolation from the various other aspects of the survey, the reader is also advised to consider *during* questionnaire design its relationship with survey objectives, population, data collection, coding and data capture, editing, imputation, confidentiality, and testing.

Since this paper is not intended to be a comprehensive discussion of either survey or questionnaire design, alert readers, depending on their own perspectives of the various aspects of a survey, may identify omissions in the principles or may wish to exclude particular principles as more appropriate to a survey component other than questionnaire design.

The basic principles of questionnaire design as presented cover the content, wording, format, and testing of questionnaires. The questionnaire has a major impact on whether or not the survey objectives are met. Unlike other major survey components such as sample design or data processing procedures, the questionnaire directly involves the respondent. Therefore, it is essential that the content, wording, and format ensure the collection of reliable, valid, and relevant information from the respondent.

The author recognizes that although some of the principles appear obvious when stated, they are usually not so in practice. Also, some of the principles are measurable; some are not.

In the principles which follow, the term *questionnaire* is consistently used to refer to the various types of forms used to obtain information. In the literature and in practice, distinctions are often made among:

(a) a questionnaire (completed by a respondent);

¹ Larry Swain, formerly of the Census and Household Survey Methods Division, Statistics Canada; currently of the Human Resources Planning Division, Public Service Commission, Ottawa, Canada K1A 0M7.

- (b) an interview schedule (completed by an interviewer);
- (c) an administrative form (completed by a respondent or an official representative of the survey organization);
- (d) a form used to record observations or measurements (completed by an official representative of the survey organization);
- (e) a form used when transcribing information from existing administrative records (completed by an official representative of the survey organization).

For simplicity, the term *questionnaire* is used herein to represent all such forms. In addition, the term *questionnaire item* is used to represent the particular question or statement requesting information, including the response categories or space for response.

The term *survey* is used generally to represent any data collection activity, including sample surveys, censuses, and administrative data collection.

2. CONTENT

1. All questionnaire items should be directly related to the objectives and uses of the survey.

It is a reasonable goal that the collection of information be designed to minimize response burden by techniques such as reducing the number of questions. Exclusion of questionnaire items only remotely related to the objectives and uses of the survey is a means of satisfying this goal.

In addition, questionnaire items that ask for irrelevant information unnecessarily contribute to the overall length of a questionnaire and may provoke suspicion in respondents, factors which may lead to increased non-response rates (a possible source of bias), to a poorer quality of data because of fatigue or lack of concentration by interviewers or respondents, and to increased costs, both financial and temporal, to the survey sponsor and to the respondents.

For the questionnaire designer, the very act of relating each questionnaire item to the survey objectives and uses helps ensure that these objectives and uses are well defined and will indeed be satisfied by the questionnaire.

2. If a questionnaire contains items that, although relevant to the survey, may not appear so to respondents, then an explanation of the reason for their inclusion should be provided to respondents.

Classification variables such as age, sex, marital status, size of organization, number of employees and variables such as name, address, and telephone number (used for follow-up procedures or for editing purposes) are possible examples where an explanation to respondents should be considered for inclusion (at least at a general level).

3. Only those questionnaire items for which responses can be provided easily and with sufficient reliability should be included.

Where information is requested through recall by respondents, the events should be sufficiently recent or familiar to the respondents; where the request can be satisfied from available records maintained by respondents, the effort (including both time and cost) required to obtain the information should not exceed the benefits to be gained by acquisition of the information.

Because of potential definitional ambiguities, increased response burden, and processing errors, it may be advisable that respondents not be asked to process information to complete a questionnaire item. It may be easier and more accurate for respondents to be asked for the specific information already available to them, to be processed later by the survey organization.

4. Respondents should not be asked questionnaire items for which they cannot be expected to provide any response.

Questionnaire items should not presume that the respondent has knowledge or awareness of a specific topic or engages in a particular activity. Filter questions can be used to exclude a respondent from a subsequent questionnaire item or sequence of items if those items are irrelevant because of the respondent's own particular characteristics, circumstances, or opinions.

Should respondents encounter many irrelevant items, they may feel that the survey questionnaire had been given to them in error. This could contribute to non-response or to poor relations with respondents.

The use of a filter question also serves to identify clearly whether or not a respondent is required to answer a subsequent questionnaire item or sequence of items. This is useful during survey processing and subsequent analysis. If a response to a questionnaire item is blank, it may be difficult to distinguish between the situation in which the reason for the blank is non-response (a refusal or an accidental omission), and that in which it is because the question does not apply (in the case of a numerical answer, the question may seem not to apply when the answer is legitimately zero). A filter question helps resolve this problem by identifying which respondents should have answered the questionnaire item.

Complex skip patterns, however, should be avoided, especially for those questionnaires completed by respondents themselves. Also, the number of filter questions should be minimized.

For items requiring a numerical answer, an alternative to a filter question is the inclusion of a *None* category.

3. WORDING

1. The phrasing of a questionnaire item should be appropriate to the respondent.

If a respondent does not understand a questionnaire item, it is probable that the response to that item will be inaccurate or not be given. Words, phraseology, and sentence structure familiar and appropriate to those providing the information should be used.

Abbreviations should be avoided unless they are understood by respondents.

2. Where there is sufficient demand, questionnaires should be translated into other languages.

Steps should be taken to ensure that the translated version corresponds adequately to the original version with respect to the intended meaning.

3. The questionnaire designer should choose the type(s) of questionnaire items most appropriate to obtain the required information while minimizing the response error and response burden in obtaining that information.

The types of questionnaire items for consideration are the open-response or free-answer type, the closed-response or fixed-answer type, and the fill-in-the-blanks type. Closed-response types are those items for which answer categories are provided. Fill-in-the-blanks types, although they appear to be open-response because no answer categories are explicitly provided,

are actually implicitly closed from the respondent's point of view in that the choice of answers is usually limited to a number, a day of the week, a province, etc.

Generally, closed-response questions entail less respondent and/or interviewer burden, since they do not require respondents to formulate and answer in their own words nor do the answers have to be recorded verbatim.

4. When a decision among two to more well-defined alternatives is required, a closed-response or fill-in-the-blanks type of questionnaire item should be used.

When all the alternatives are too numerous to be listed, then the use of the category *other* to represent a number of infrequently occurring responses, the use of an open-response or fill-in-the-blanks type of questionnaire item, or the collapsing of alternatives into fewer categories is recommended. In fact, it may be appropriate to use a fill-in-the-blanks type, where the response categories and numerical codes are included in a separate instruction booklet accompanying the questionnaire. In business, agriculture and institutional surveys, fill-in-the-blanks types of items are frequently used for questions that require a numerical response. The choice and number of categories in a closed-response type depends on the complexity of interpretation of the concept, the uses to which the data will be put, and the prior information available to the questionnaire designer.

5. When the alternatives to a question are not well-defined, an open-response type of questionnaire item should be used.

Open-response types of questionnaire items are frequently used in preliminary research or exploratory studies to generate specific hypotheses and to structure items for subsequent questionnaires. The open-response type of questionnaire item may also be used as a means of probing for additional or qualifying information, for purposes of verification of other questionnaire items, for use in interpreting data, as a change of pace, or as an introduction to a new topic.

6. If ease, timeliness, and cost of processing the data for capture are important considerations, closed-response types of questionnaire items should be used.

Open-response types of questionnaire items require coding of the information provided, an operation which can be both costly and time-consuming and is also subject to errors of interpretation and procedure.

In addition, with open-response types of questionnaire items, no specific frame of reference is provided, leading to the choice of varying frames of reference on the part of respondents. These varying frames of reference and the provision of varying amounts of information by respondents cause difficulty in the recording, coding and analysis of responses. On the other hand, a closed response provides a specific frame of reference, which although avoiding the above problems, may artificially induce a response. This is especially true when the respondent has little or no information or opinion about a particular topic. The questionnaire designer must therefore be aware of the possible frames of reference of respondents before choosing a type of questionnaire item.

Once the type and wording of a questionnaire item have been decided upon, restrictions are placed on the uses to which the information can be put, the specific hypotheses which can be tested and the analyses that will be applied to the item. This implies that the determination of objectives, uses, hypothesis testing and analyses is a prerequisite to the final version of the item. This determination does not preclude that the data will suggest additional analyses and uses within the limits imposed by the questionnaire items themselves.

In addition to the above considerations, the past experience of the questionnaire designer will contribute to the choice of suitable type(s) of questionnaire items in particular situations.

7. Response categories for closed-response types of questionnaire items should be non-overlapping and exhaustive (that is, mutually exclusive and comprehensive).

The response categories of a particular questionnaire item should be distinct and include all possibilities.

The distinctiveness of response categories does not preclude the applicability of more than one response to a particular questionnaire item. In such a case, a note such as *check as many as apply* should be included as part of the question.

Where response categories are such that only one response is to be provided to a particular questionnaire item, a note such as *check one item only* should be included as part of the question (except in the most obvious cases, for example, where the response categories are *Yes* and *No*). In those cases where more than one response can be applicable but where the designer wishes that only one item be checked in order to restrict responses, a note such as *check the most appropriate item* should be provided.

8. The units of response should be specified.

Either the units of response (e.g., kilograms, tons, per cent, hours per week) should be included in the questionnaire item or the respondent should be asked to specify them. Otherwise, there may be ambiguity as to which units were actually used.

9. Standardized concepts and definitions should be used.

To facilitate comparison of survey data with other sources of information (publications, other surveys) and to maximize the usefulness of the data (including secondary analysis), standardized (commonly understood and used) definitions should be used where they exist, and are well-defined, appropriate and up-to-date. Statistics Canada publishes standards related to occupational classes, industrial classes, commodities, geography and specific social concepts. In addition, Census concepts and categories are frequently used as standards.

10. The wording of questionnaire items should be specific, definitive, consistent, brief, simple and self-explanatory.

Survey concepts and terms that are new to respondents or subject to misinterpretation should be explained, defined, or avoided. To ensure consistent interpretation, the proper frame of reference (e.g., time reference, location, category of expenditure) should be provided. If consistency is required (e.g., different time references for different items), the change should be highlighted in the questionnaire.

Where several words can be used interchangeably, one of these should be selected and used throughout the questionnaire. If a synonym of a word already encountered is used in its place, respondents and others may assume that a different meaning is intended.

11. Double-barreled questions should be avoided.

A double-barreled question allows the respondent to make only one response although it is actually two questions in one. From the response, it is not possible to discern which

of the two ideas was answered or whether both were answered. The two issues should be asked separately except in specific circumstances where two issues necessarily have to be asked together to convey the proper meaning. In such a case, it should be made clear to respondents that the two issues are both to be considered together.

12. Leading questionnaire items should be avoided.

A leading questionnaire item is one that is worded or formatted in such a way as to induce a respondent to choose a particular alternative or set of alternatives.

Some questions can be considered to be leading if they present options that may be perceived by respondents as socially unacceptable without an assurance that the respondent is made to feel that there would be no stigma attached to their response.

In attitudinal surveys, two basic principles have evolved to reduce (but not necessarily eliminate) response bias. The distribution of alternative answers should balance to provide approximately as many positive answers as negative to avoid leading respondents in one direction. Secondly, where there exists a series of items that have the same response alternatives, the sequence of items should either contain a mixture of positive and negative statements, be broken up, or be presented in a varied order to reduce the incidence of respondents answering in the same manner throughout the sequence (even though it may be inappropriate), without thinking very carefully about the particular responses.

4. FORMAT

1. Every questionnaire or questionnaire package should contain explanatory introductory material.
2. The introductory material should state the title of the survey, the name(s) of the sponsoring institution(s), and the purposes(s) of the survey.
3. An assurance to respondents of the confidentiality of the data that they provide should be considered.
4. The name (if appropriate) and telephone number or postal address of a contact within the sponsoring institution(s) should be included on the questionnaire in order that respondents may obtain additional information related to the survey, should they require it.

Generally, the introductory material may be in the form of a letter or brochure sent to the respondent; it might be a prepared statement made by an interviewer; or it can appear on the questionnaire itself. The introduction contains essential background information to respondents for the purposes of identification, legitimacy and notification of legal rights (if applicable).

5. Suitable identification should appear on the questionnaire.

For the purposes of estimation, field control, linkage with other records or follow-up on non-respondents, appropriate identification (numerical or otherwise) should be included on the questionnaire.

6. Questionnaire items and pages should be numbered.

To facilitate administration by interviewers, completion by respondents, and coding operations and instructions, questionnaire items and pages should be numbered consecutively (using either letters or numbers) throughout the questionnaire. If questions are written on

both sides of a page, an instruction (e.g., *over*) should appear at the bottom of the first side to ensure that the questions on the second side are completed.

7. The print on the questionnaire should be such that it can be easily read by the average respondent.

The person completing the questionnaire must be considered when determining the size of the type face (for example, small print could cause problems for those with poor eyesight) and the colours and contrasts of paper and type to be used. It is usually advisable to have different type face (size or type of characters) used for questions and instructions so that they can be easily distinguished.

8. Instructions for completion should be included on or with the questionnaire.

To help ensure that the questionnaire is completed properly by respondents, interviewers or other officials, brief but clear instructions should appear on or with the questionnaire (e.g., in an interviewers' manual or an instruction manual). However, questionnaire items should be as self-explanatory as possible to avoid complex sets of instructions.

For questionnaires being read by an optical character reader, clear instructions should be provided to help ensure their proper completion.

Instructions to respondents or interviewers for skipping items following filter questions should be sufficiently obvious and easy to follow. The use of arrows and directions may be appropriate. Complex skip patterns should be avoided, especially for questionnaires completed by respondents themselves.

9. The instructions for return procedures should be included on the questionnaire.

For a questionnaire which is to be returned by mail, the name and address of the person (or organization) to whom it is to be returned should be included on the questionnaire itself. Introductory letters and return envelopes can easily be mislaid or separated from the main body of the questionnaire.

The deadline by which respondents are to return completed questionnaires should also be stated.

For a questionnaire which is to be picked up by a field representative, space for the name and telephone number or postal address where the representative can be contacted and the date and approximate time of pick-up should be included on the questionnaire.

10. The numerical fields and codes used for data capture purposes should appear on the questionnaire (when capture is to be directly from the questionnaire).

When appropriate, data may be captured more quickly with fewer errors directly from the questionnaire itself. In such a case, the numerical fields and codes should be easily read by those performing the data capture but should not be a distraction to the respondent, interviewer or other official completing the questionnaire.

When data are to be coded before data capture, the coding boxes may appear on the questionnaire or on a separate sheet. When coding boxes do appear on the questionnaire, they should be clearly distinguished from answer boxes, perhaps with the *Office Use Only* designation or through appropriate shading.

Coding and data capture are often considered as steps that follow questionnaire design. It is essential for efficient implementation that they be considered during questionnaire design.

11. The format of answer spaces should be consistent throughout the questionnaire, with sufficient spacing for purposes of readability and accommodation of the responses to the questionnaire items.

Consistency of layout for response facilitates the task of a respondent, interviewer or official and aids in reducing error caused by inadvertent omission of a questionnaire item, an incorrect response, or a transposition of responses.

It may be useful to use different shapes for *check-off* type answers and numerical answers. One convention sometimes used is circles for the former and boxes for the latter.

There should be generous spacing on the questionnaire: to facilitate administration; to make the questionnaire more attractive and readable; and to provide the respondent, interviewer or official with sufficient space for the response to the questionnaire item.

12. Questionnaire items should be sequenced in a logical order for ease of completion and to provide the proper frame of reference.

The sequence of questionnaire items should appear logical to the respondent (a logic that may be different from that of the questionnaire designer), with questionnaire items related to one another grouped together. One sometimes recommended method is to have questions proceed from the most general questions to the most specific. Question ordering should try to anticipate the order in which respondents will supply information. The questionnaire designer should recognize that a question may prompt an answer not only to that question but also to another question which (hopefully) follows very shortly.

Transitions between sections of questions should be smooth. Section headings or introductory statements to sections should be used. For questionnaires used in transcription from other documents, a logical sequence would be that of the source document.

In attitudinal surveys, the questionnaire designer should avoid conditioning respondents in the early questioning to a frame of reference which could bias responses to later questions. For example, questionnaire items regarding the awareness of a concept should precede any other mention of that concept. Sensitive questions should be placed within the context of related questions so as to justify their inclusion as much as possible and desensitize the questions somewhat.

13. The final version of the questionnaire should contain no typographical or grammatical errors.

The inclusion of errors on the questionnaire may have an adverse effect on data quality in that the questionnaire may not be treated seriously or may be misunderstood by those completing it. In addition, errors may contribute negatively to the image of the survey organization in the eyes of the public.

5. TESTING

1. Questionnaires administered for the first time or containing substantial modifications should be tested prior to their use as a collection document.

Just because all principles described in the previous principles have been followed, there is no guarantee that the proposed questionnaire will fully satisfy the objectives of the survey no matter how conscientious the researcher has been in designing the questionnaire. There are almost always unforeseen problems that occur in the administration of a questionnaire.

As a result, it is essential that a pretest of the questionnaire be implemented for all new surveys and for already existing surveys on which substantial modifications have been made in order to determine whether the objectives are likely to be met by the proposed questionnaire.

Some aspects of the questionnaire that the designer may test are the following: the wording, sequence and layout of the questionnaire to determine whether the questions and their flow are understood by respondents and interviewers; the necessity for inclusion of particular questions; the choice of types of questions; the use of specialized questioning techniques such as ranking or rating questions; the structure and definition of response categories; the degree of usage of the "other" category in questions; the ease of administration of the questionnaire; the time to administer various sections of the questionnaire; translation of the questionnaire; the possibility of bias in the questions; the nature of ethnic, regional or linguistic differences; the reasonableness of the questionnaire with respect to its demands on the respondent; the suitability of the questionnaire for measuring the concepts on which measurement is required; letters of introduction or introductory procedures; and the suitability of the method of collection.

A pretest should be done on at least a small sample of respondents (usually twenty to thirty) from the target population. It is preferable that the respondents be selected from the various subpopulations of the target population where differences or problems are likely to occur. Possible variables for definition of the test subpopulations are geographic region, educational background, age, sex, language, size of firm and type of industry. Depending on the particular purposes of the pretest, either a probability or a non-probability sampling scheme may be required for the selection of respondents, although in most cases, the latter is employed. One possibility is to use a focus group discussion of the questionnaire as a part of the pretest procedure.

The method of collection used for the pretest should be identical to that planned for the main survey. However, a personal interview is recommended for at least a portion of the pretest respondents so that the interviewer can then record the respondents' reactions, both verbal and non-verbal, as well as their own suggestions and impressions. After each test interview, the interviewer can discuss difficulties that the respondent had, the interpretation of questions and response categories, and so on. These difficulties can then be discussed with the designer of the questionnaire, for example, in the context of a meeting among the questionnaire designer and the pretest interviewers to debrief them on the interviews. For some pretests, it may be preferable to use experienced, skilled interviewers in order to maximize the usefulness of the pretest.

The pretest is an often-neglected procedure. It will almost always suggest improvements or will at least give the designer some assurance that the questionnaire used in the main survey, a much more expensive proposition, will likely proceed fairly efficiently. Of course, there is never any guarantee that all problems will be solved, but most major ones should be. A pretest need not be expensive and need not require a great deal of time for implementation and is recommended for all new or modified questionnaires.

ACKNOWLEDGEMENTS

The author wishes to thank not only the referee and editorial board for their helpful comments but also the many persons who responded to an earlier distribution as part of a partial draft of "Survey Design Standards and Guidelines".

REFERENCES

- ANDERSON, J.F., and BERDIE, D.R. (1974). *Questionnaires: Design and Use*. Metuchen, N.J.: The Scarecrow Press, Inc.
- BERTHIER, N., and F. (1971). *Le sondage d'opinion*. Paris: Bordas.
- BON, F. (1974). *Les sondages - peuvent-ils se tromper?* France: Calmann-Lévy.
- CARSON, E. (1974). Questionnaire Design, Some Principles and Related Topics. Unpublished Manuscript, Statistics Canada.
- CORBIN, R., SWAIN, L., and WILHELM, E. (1977). Exposé pour un atelier sur la conception des questionnaires. Unpublished manuscript. Statistics Canada.
- CORBIN, R., SWAIN, L., and WILHELM, E. (1977). Outline for a Workshop on Basic Questionnaire Design. Unpublished manuscript, Statistics Canada.
- GHIGLIONE, R., and MATALON, B. (1978). *Les enquêtes sociologiques: théories et pratique*. Paris: Colin.
- JAVEAU, C. (1974). *L'enquête par questionnaire. Manuel à l'usage du praticien*, third edition. Bruxelles: Institut de Sociologie de l'Université Libre de Bruxelles.
- MOSER, C.A., and KALTON, G.J. (1972). *Survey Methods in Social Investigation*, second edition. New York: Basic Books.
- OPPENHEIM, A.N. (1966). *Questionnaire Design and Attitude Measurement*. London: Heinemann.
- PAYNE, S.L. (1951). *The Art of Asking Questions*. Princeton: Princeton University Press.
- STATISTICS CANADA (1979). Basic Questionnaire Design, second edition. Unpublished manuscript, Statistics Canada.
- STATISTICS CANADA (1981). Conception des questionnaires, Manuel d'atelier, third edition. Unpublished manuscript, Statistics Canada.
- WARWICK, D.P., and LININGER, C.A. (1975). *The Sample Survey: Theory and Practice*. New York: McGraw-Hill.

An Overview of the Strengths and Weaknesses of the Selected Administrative Data Files¹

RAVI B.P. VERMA and PIERRE PARENT²

ABSTRACT

Twelve administrative data files are reviewed to determine if some of them could be used to derive migration data, in case the universality of the currently used family allowance files be limited, as a result of federal legislation.

It is found that none of the twelve files have strengths and weaknesses strictly comparable to those of the family allowance files. Further developments of the Health Care, and to a lesser extent the Old Age Security files are highly recommended.

KEY WORDS: Administrative files; migration; qualitative evaluation.

1. INTRODUCTION

In Canada, both family allowance and income tax files have a wide range of utility in producing the migration and population estimates for the different geographic areas (see Statistics Canada Catalogue Nos. 91-001, 91-210, 91-211 and 91-212). Data from the family allowance files are made available within 2 to 3 months after the reference date. In contrast, income tax data are available within 12 to 15 months after the reference date. However, income tax data provide the estimates of migration flows for the census divisions, and also by age and sex.

In terms of accuracy of population estimates, both family allowance and income tax files are good and they are comparable (see Norris and Standish 1983; Norris 1983; Verma *et al.* 1984; Verma and Basavarajappa 1985). One of the special features of the family allowance and income tax files is the fact that they are national in character. Another feature is that the records contain addresses with the postal codes. Thus, this could provide the migration information for local areas. However, in recent years, there seems to be some possibility that family allowance could cease to be universal as a result of government legislation. For example, coverage might be limited to the lower- and middle-sectors of the population. If this file ceased to be universal, its utility as a migration data source would be very severely limited. Hence, our population estimation activities would be jeopardized which in turn would affect other programs as revenue sharing, involving the annual distribution of \$20 billion among provinces.

For this reason, alternate sources need to be explored. An attempt is made here to assess the strengths and weaknesses of some of the selected administrative data files for estimating migration and population for provinces and territories, census divisions, census metropolitan areas and other regions in Canada.

The twelve administrative files are qualitatively evaluated as an alternative to family allowance files. On the basis of their strengths and weaknesses, they are divided into the following three groups:

¹ Abridged version of the paper presented at the meetings of the Federal-Provincial Committee on Demography held on November 28-29, 1985, Ottawa, Canada.

² Ravi B.P. Verma and Pierre Parent, Demography Division, Census and Demographic Statistics Branch, Statistics Canada, 4th floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Major potential files for estimating migration flows

- i) Health Insurance Files
- ii) Old Age Security File

Major potential files used as a symptomatic indicator of population change and net migration

- iii) Hydro Connections
- iv) Telephone Customers
- v) School Enrollments

Other files with limited or uncertain potential for estimating migration flow/net migration

- vi) Driver's License
- vii) Building Permits
- viii) Unemployment Insurance Beneficiaries
- ix) Labor Force Survey
- x) Voters' List
- xi) Retail Sales
- xii) Trucking Statistics

1.1 Criteria for Evaluating Administrative Data Files

The assessment of the usefulness of the various administrative data sources for estimating interprovincial and intraprovincial migration is done with respect to ten criteria: universe, coverage, method of determining migration information, types of migration, characteristics of records, reference date/period (and monthly availability), time-lag, historical availability, consistency and computerization (Almond 1982).

The new data source would have high potential if it contains features of the family allowance files, as described in Table 1. The most important criteria are: coverage, timeliness, consistency, monthly or quarterly availability, disaggregation using the postal code or other geocodes. The file or set of files that can meet these standards would probably qualify as replacement source to family allowance.

2. MAJOR POTENTIAL FILES FOR ESTIMATING MIGRATION FLOWS

Health Insurance and Old Age Security files are major potential files for estimating migration flows among provinces, territories and census divisions. Strengths and weaknesses of each of these two files are presented below.

2.1 Health Insurance File

Health Insurance is a provincial responsibility. Each province thus keeps a file of people eligible for the program. All residents in the province (including newly arrived immigrants and foreign students) are covered by the provincial insurance, except for RCMP and Armed Forces personnel, and for the federal penitentiary inmates, covered by the federal government. Everybody who establishes its residence in a province must fill out a proper application, from which data on in-migrants, by province of origin, and on international immigrants can be compiled. Virtually complete coverage, monthly availability, minimal time lags and information usually detailed by age, sex and family composition of the migrants are the main strengths of the files. There should also be a very strong incentive for interprovincial migrants to apply to the program. Consequently, migration data should be reliable.

Table 1
Description of the Administrative Data Files Currently
Used to Derive Migration Data in Canada

Criteria	F.A. Monthly Statistics Report	F.A. M0024 File
Universe	Children in payment of F.A.	Children entitled to F.A. (as opposed to "in payment")
Coverage	25% of total population in 1984. Virtually 100% of children aged 0-17	Similar to F.A. Monthly Statistics
Method for determining status	Compilation of change of address notices	Compilation of change of address notices
Types of migration	Interprovincial migration, by province of origin and destination	Similar to F.A. Monthly Statistics, plus international migration
Characteristics	Origin-destination. Age: total 0-17 only. Family size: refers to the number of children in family	Origin-destination. Age: year and month of birth. Language (E or F). Type of account (regular, foster, foreign or agency)
Reference date/period	Month: refers to the amount of information processed during that time	Month: refers to month of real migration
Time-lags	Data processed a given month is available at the end of that month and refers to migration of approximately two months earlier	Data released semi-annually. Contains information on last six months' migration. Available approximately 3 months after end of semi-annual version
Historical availability	January 1974 onwards for children migration data. From 1947 to 1973, only information on family migration was available	December 1977 to present
Consistency	OVER TIME: change in 1974. Slight, problems since 1980. AMONG PROVINCES: good	OVER TIME: generally good. Slight problems since 1982. AMONG PROVINCES: problems with Ont. since 1982. Slight problems with Nfld. and N.S. in 1983
Level of computerization	In provincial offices, yes. But data are sent to Health and Welfare Canada central office on print-outs	Yes, well developed

Note: F.A. is an abbreviation for Family Allowance.

Table 1

Description of the Administrative Data Files Currently
Used to Derive Migration Data in Canada (Concluded)

Criteria	F55 Program	Revenue Canada File
Universe	Children entitled to F.A.	Tax filers (must have filed two consecutive years)
Coverage	Similar to F.A. Monthly Statistics	Filers matched two consecutive years total up to approximately 75% of population aged 18 and over
Method for determining status	Symptomatic indicator	Comparison of the return address of matched returns. Correction is brought for unmatched returns
Types of migration	Net migration	Intraprovincial, Interprovincial and International
Characteristics	Number of children by geographical area. Age	Origin-destination. Broad age-sex group
Reference date/period	Twice a year, as of June 1, and December 1 (refers to the number of children entitled to F.A. as of these dates)	Year: refers to the period between two consecutive filings, i.e. approximately the April-March period. Used as June-May data
Time-lags	Available approximately three months after reference date	Preliminary available 6-8 months after end of reference period. Final data, 10-12 months
Historical availability	December 1977 to present, with entitlement information. Available back to 1974 for children in payment	1966-67 to present
Consistency	Generally good	Changes in tax laws results in change in coverage and in number of matched returns over time and provinces
Level of computerization	Yes, well developed	Yes, well developed

There are also, however, certain weaknesses. The fundamental limitation is that neither Ontario nor Quebec can provide migration data. In the latter case, however, new developments are promising, but for Ontario, nothing is expected. Unless a special source is derived for Ontario, this would compromise the high potential of this file. There could also be a consistency problem, since each province independently administers its file.

At the subprovincial level, migration could also be derived since the Provincial Health Care offices should be informed of any change of address. In the facts, however, all changes are not known.

Health Care files could also be used in regression estimates, especially in provinces that run periodic address checks to clean the file and count only the desired population.

2.2 Old Age Security Records

Health and Welfare Canada is responsible for the administration of the Old Age Security file. Canadian residents aged 65 and over who totalled a sufficient number of years of residence in the country are eligible. It represents approximately 10% of the total population. Coverage among eligible people is virtually universal. Also the financial incentive to report change of address is very strong. Another strength of the file is its timely availability. Information on people moving in a given month is compiled and received by Statistics Canada two or three months later. Finally, Old Age Security, being a federal program, provides comparable data for the provinces; even if the information is compiled by provincial regional offices, they all follow the same procedure.

The main shortcoming of this file for migration estimates purpose is the fact that it refers to a small portion of the population (varying from 7.3% in Alberta to 12.2% in P.E.I.), the elderly moreover showing a rather different migration pattern than the rest of the population. Unlike child migration, which can obviously be related to adult migration and then be blown up to estimate total migration, no similar efficient method could be developed to estimate total migration from the Old Age Security file. Although this could not be used as the main source for migration estimates, however, this file could provide a very interesting estimate of the elderly migration.

3. MAJOR FILES USED AS SYMPTOMATIC INDICATORS OF POPULATION CHANGE AND OF NET MIGRATION

Data from some administrative files could be useful for generating total population estimates. For example, School Enrolments, Hydro Connections or Telephone Residential Customers could be used in regression techniques as symptomatic indicators (see McRae 1985 for an application of Hydro Connections to population estimates). This method and the corresponding sources are generally used for producing small area population estimates, but if no other technique gives valuable estimates at the provincial level, these sources will be seriously considered.

3.1 Hydro Connections

Electric companies keep files of their customers. Information on the type of account (residential, commercial, farm, ...) and the address and postal code of the customer are available. Coverage of residential households is virtually complete. Sometimes there is only one file for the province, but sometimes 2 companies (Manitoba and Newfoundland) or even more (B.C. and Ontario) provide the electric facilities within the province. In most provinces data can be produced for the entire territory, as of any date and within a short time-lag, but for a few provinces it can be hard to get the data. The main weaknesses of the files are of two kinds. In addition to the previously cited problem there may also be slight inconsistencies due to the difference between provincial definitions of residential households (since

it responds to administrative criteria), and even within one province, if more than one company is involved. Nevertheless, Hydro Connections could be a very good source for population estimates. As a matter of fact, they were tested in British Columbia, where population estimates for municipalities and school districts were produced. The results were good. This method could also be tested and eventually be extended to provincial level estimates, if need be.

3.2 Telephone Companies

In Canada, telephone services are insured by 14 major telephone companies. Information on customers with residential lines (address and postal code) is available. The situation is roughly similar to that of the Hydro Connections files. Data can usually be obtained for specified dates within a rather small delay and the coverage is fairly high. Here again, more than one company may serve a given province, and also, a company may serve more than one province. Despite the fact no estimate based on Telephone files has been tested in Statistics Canada, it is felt that they have the potential to produce good results.

3.3 School Enrollments

Each provincial government maintains a computerized file on students enrolled in its primary and secondary school system, containing information on school addresses with the postal code and on the number of students, by age and grade. Information on the number of students refers to September 30 and is available between 4 and 10 months after the reference date, the time-lag varying by province. The coverage of students is also very good.

There are some weaknesses associated with this file. For example, its annual character plays against its use for producing quarterly estimates. Also its date of reference (September 30 instead of June 1), along with the up to 10 months delay is another handicap. At the subprovincial level, finally, it often can be observed that some students reside in a given administrative region, but go to school in a different one. This also could affect the quality of the estimates. It should be pointed out here that the school enrolment data, at one time, were used in Statistics Canada (and also by the U.S. Bureau of the Census, using a component method developed by them. See U.S. Bureau of the Census 1973, Chap. 23, p. 51); the deviations associated with that method were much higher than those with other methods. In case no other file could provide adequate population estimates, regression estimates with that file could produce acceptable results, at the provincial level at least.

4. FILES WITH LIMITED POTENTIAL

4.1 Driver's License

Each province maintains a file listing persons aged 15 (or 16, or 17) and over licensed to operate a motor vehicle. Using the provincial files, migration could be estimated in two ways: 1) compilation of changes of driver's address for estimating flows of migration; and 2) as a symptomatic indicator of the population change, through the variation of the number of people licensed in a given region. Currently, Ontario uses drivers' licenses to estimate intraprovincial migration, but very few other provinces could provide migration flow information, especially at the subprovincial level. In order to do so, it would require too much work and consultation with the provincial ministry. Despite the fact that drivers are forced by the law to report their change of address, not all do so, and no sufficiently detailed statistics are available.

The driver's license file could also be used in regression techniques. Data available at any specified date in many cases and short delays are positive points. However, coverage and consistency concerns might affect the quality of the data. For example, 83% of adults in Saskatchewan own a driver's license, as against 73% in Manitoba, 85% of males and 62% of females accounting for the latter province's average proportion. In addition, the poor, comparatively recent immigrants, and Indians and residents of remote communities in the

North have below average rates for holding licenses (Stock 1981, p. 44). For estimate purposes, it is often preferable to have a 100% coverage of a small subpopulation (e.g. children) than an 80-85% coverage of a large subpopulation (e.g. adults), especially if the coverage is selective with respect to migration. Although it does not necessarily make a file inappropriate for estimate purposes, it affects its potential.

4.2 Building Permits

Statistics Canada collects new building permits for cities and rural areas in Canada. On average, the coverage rates vary between the urban (98.5%) and the rural (62.5%) areas. The building permit data are available on a monthly basis at the census division level. These data could be also used as a symptomatic indicator of the population change. However, one of the weaknesses of the building permit data is the fact that they refer to the date of permit. Due to this, it is not certain whether the building has been constructed and also, whether it has been occupied. Another weakness is the fact that the number of permits issued is not necessarily directly related to population change, especially in the case of a decreasing population.

Thus, the use of building permit data also seems to be limited in estimating population for the different geographic areas.

4.3 Unemployment Insurance Commission

The Unemployment Insurance Commission keeps a list of the beneficiaries of the program. A 10% sample of this file has been developed to produce statistics and it could provide migration information. However, this file could hardly be used to estimate migration in Canada. First, a 10% sample of unemployed corresponds to less than 1% of the population. From such a small subpopulation, flows of migrants between provinces could not be derived. Also the non-representativity of that sample (young adults representing a good part of non-employed) calls for suspicion concerning the migration data from that file.

The Commission also maintains a file of wage earners who are contributing to the Unemployment Insurance program. However, no in depth analysis of this file has been done.

4.4 Labour Force Survey

In 1982, Statistics Canada conducted a sample survey of 56,000 households in Canada. The civilian non-institutionalized population aged 15 and over, included in the sample, residing in all provinces were asked a question on their migration history of the past 5 or 6 years. Other valuable information is also available. However, its very small sample (approx. 1/2% of the population) and the fact that the survey was conducted only once eliminates the Labour Force Survey as migration estimates source.

4.5 Voters' List

Data on voters are generally available in Canada. Federal and provincial election lists could easily be obtained while obtaining municipal lists would necessitate more work. Those lists give information on the number of canadian citizens aged 18+ (landed immigrants are included at the municipal level only). They cover an average 90-95% of the target population. The main shortcoming of that source is that it is not available at regular intervals. Federal and provincial lists are made for elections about every 4 years at dates that are not useful for estimation purposes. It thus seems pointless to consider voters lists.

4.6 Retail Sales

Data on retail sales are collected by Statistics Canada on the basis of sales figures from large stores and from a sample of smaller businesses. These data are collected on a monthly basis and they are made available 3 months after the reference date. These data could be used as a symptomatic indicator of the population change. However, the utility of this data

set seems to be limited in the case of population and migration estimations. This could be due to the fact that retail sales are heavily affected by the economic fluctuations which may not accurately reflect changes in the size of population.

4.7 Trucking Statistics (Moving Companies)

Statistics on a sample of five major moving companies are available in Canada. They cover about 90% of all moves. The interprovincial migration flow could be assessed by weighing the number of reported moves between two different provinces/territories. However, trucking statistics are seriously affected by a time-lag of two years or more.

5. CONCLUDING REMARKS

In this report, an overview of strengths and weaknesses of twelve administrative data files has been presented in order to make recommendations for selecting an alternative data source to the family allowance files. It has been found that there is no file with strengths and weaknesses strictly comparable to those of the family allowance files. However, if the family allowance files cease to be universal, one could suggest the following recommendations:

- Continue further developments in the use of the provincial health insurance file and the Old Age Security records of the federal government in order to produce the total population and migration estimates on a quarterly basis;
- Examine the quality of annual population estimates for the provinces and territories, produced by the Component Method II using the migration estimates from the provincial school enrollment data files; and
- Test the accuracy of the provincial administrative data files (health insurance files, hydro connections, telephone companies and driver's licence) as symptomatic indicators of the population change and the residual net migrants for sub-provincial areas (census divisions and census metropolitan areas in Canada).

REFERENCES

- ALMOND, M.M. (1982). An inventory of sources of Canadian migration data. Working Paper, Demography Division, Statistics Canada.
- McRAE, D.G. (1985). Use of hydro accounts in the regression population estimates model in British Columbia. Presented at the Federal-Provincial Committee on Demography, Ottawa, Canada.
- NORRIS, D.A., and STANDISH, L.D. (1983). A technical report on the development of migration data from taxation records. Technical Report, Administrative Data Development Division, Statistics Canada.
- NORRIS, D.A. (1983). New sources of Canadian small area migration data. *Review of Public Data Use*, 11-25.
- STATISTICS CANADA (Quarterly). *Estimates of Population for Canada, Provinces and Territories*. Catalogue 91-001, Ottawa: Minister of Supply and Services Canada.
- STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population by Marital Status, Age, Sex and Components of Growth for Canada, Provinces and Territories*. Catalogue 91-210, Ottawa: Minister of Supply and Services Canada.
- STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population for Census Divisions and Census Metropolitan Areas (Regression Method)*. Catalogue 91-211, Ottawa: Minister of Supply and Services Canada.

- STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population for Census Divisions and Census Metropolitan Areas (Component Method)*. Catalogue 91-212, Ottawa: Minister of Supply and Services Canada.
- STOCK, R. (1981). *Migration Estimates from Current Administrative Files: Data Sources and Methodologies*. Canadian Plains Research Center, University of Regina.
- U.S. BUREAU OF THE CENSUS (1973). *The Methods and Materials of Demography*. Washington, D.C.: U.S. Government Printing Office.
- VERMA, R.B.P., BASAVARAJAPPA, K.G., BENDER, R.K. (1984). The regression estimates of population for sub-provincial areas in Canada. *Survey Methodology*, 9, 219-240.
- VERMA, R.B.P., BASAVARAJAPPA, K.G. (1985). Recent developments in the estimation of population for small areas in Canada by regression. Presented at the International Symposium on Small Area Statistics, Ottawa, Canada.

Use of Administrative Data Files for Migration Estimates: A Case Study of Driver's Licence File in Ontario¹

RAGHUBAR D. SHARMA and CHEUK WONG²

ABSTRACT

In Canada, provincial and federal demographers have attempted to use various sets of administrative data to estimate migration flows. This paper presents the development of intra-provincial migration estimates using driver's licence data in Ontario. An evaluation of these migration estimates has been carried out by comparing with those derived from the income tax data by Statistics Canada. Both files provide equally good and complimentary estimates of intra-provincial migration.

KEY WORDS: Administrative files; Population estimates; Component method; Small areas; Error of closure; Intraprovincial migration.

1. INTRODUCTION

Migration is an important component of population projections, and population estimates. As no records regarding the movement of population are kept in Canada, demographers in the federal and provincial governments have attempted to use various sets of administrative data to estimate migration flows. Statistics Canada uses revenue data (Norris and Standish 1983), British Columbia utilizes hydro-hookups (McRae 1985), and Alberta uses health care records (Alberta Bureau Statistics 1985). Since 1979, Ontario has been using drivers' licence address changes to estimate intra-provincial migration. Apart from the quality aspect, one major attractiveness of the driver licence data is in its timeliness. There is only a 4 to 5 week time lapse between receiving the data and the date of reference compared with over one and one-half years in revenue data. In this paper we shall present an evaluation of estimates of intra-provincial migration derived from the driver's licence data in Ontario. In the U.S.A., the State of California also uses driver's licence address changes for the estimation of intra-provincial migration (Hoag 1984).

2. DRIVER'S LICENCE DATA FILE

Information on driver's licence address changes is made available by the Ontario Ministry of Transportation and Communications (MTC). A driver is required to notify the Ontario Ministry of Transportation and Communications within 90 days of his/her change of address. The information is available at the postal code area level. These postal code areas can be converted into such subprovincial areas as, counties, regions and municipalities. As Table 1 indicates, data are available for the past seven years. Since 1979, data are also available for each quarter of these years.

More than a million changes of addresses are recorded every year. The majority of these moves tend to be within census divisions (that is, county or regional municipality). However, net inter-county movers averaged only about 22,000 per year. Table 1 indicates that about one-third of the records do not provide a postal code for either origin and/or destination of the mover.

¹ Abridged version of a paper presented at the meetings of The Federal-Provincial Committee on Demography, November 28-29, 1985, Statistics Canada, Ottawa.

² Raghubar D. Sharma and Cheuk Wong, Sectoral and Regional Policy Branch, Ontario Ministry of Treasury and Economics, Queen's Park, Toronto, Ontario M7A 1Y9.

In Ontario, a person becomes eligible to hold a driver's licence at the age of 16 years. More than 75 per cent of the eligible population holds a driver's licence. The elderly population and female population have a much lower tendency to hold a drivers' licence (Table 2).

3. CONVERSION OF DRIVERS TO MIGRANTS

An adjustment factor is applied to the number of drivers to arrive at the number of movers. This adjustment factor (F) is calculated as follows:

$$FA = \frac{\text{Known and Unknown Movements}}{\text{Known Movements}}$$

$$FB = \frac{\text{Total Population}}{\text{Population with a Licence}}$$

$$F = FA \times FB.$$

Table 1

Number of Total Movers and Number of Movers with
Unstated Origin and/or Destination, Ontario, 1975-1985

Year	No. of Known Movers (Inter & Intra Country)	No. of Origin and/or Destination Unstated	Total	% Unstated
1979 (Calendar Year)	881,000	0	881,000	0
1979/80	586,000	301,000	887,000	34
1980/81	566,000	306,000	872,000	35
1981/82	617,000	270,000	887,000	30
1982/83	648,000	259,000	907,000	29
1983/84	822,000	320,000	1,142,000	28
1984/85	831,000	330,000	1,161,000	28

Source: Ontario Ministry of Transportation and Communications.

Table 2

Percent of Population Holding Driver's Licence, Ontario, 1981

Age	% of Population Holding A Driver's Licence, 1981		
	Male	Female	Total
16-19	63.9	36.5	49.1
20-24	92.7	73.0	85.7
25-34	98.6	81.6	90.0
35-44	99.7	79.9	90.0
45-54	96.7	67.8	82.4
55-64	93.2	56.7	74.2
65 +	73.1	27.4	46.4
Total	90.6	62.2	75.8

Source: Ontario Ministry of Transport and Communications.

FA accounts for the unstated origins and/or destinations and *FB* accounts for non-driver's licence holders. The factor assumes that migration patterns of those who do not hold driver's licence do not differ from those who hold driver's licence. Similarly, it assumes that migration patterns of those with unstated movements do not differ from those whose movements are stated.

4. INTRA-PROVINCIAL MIGRATION ESTIMATES: DRIVER'S LICENCE VERSUS INCOME TAX FILES

Statistics Canada uses change of address as provided by a taxpayer on his annual income tax return. The number of children are estimated from the number of dependents claimed by the taxpayer. Like the driver's licence data, adjustment factors have to be introduced to the revenue data to overcome unstated postal code and people who do not file an income tax. Furthermore some taxpayers use a non-residential mailing address in their return.

The relative accuracy of migration estimates derived from the income tax file and driver's licence file needs to be tested. Three measures have been applied to test this relative accuracy of the two data sets. They are:

- A. Errors of Closure
- B. Growth Rates Test
- C. Index of Dissimilarity

Ideally, errors of closure and growth rates should be calculated from the population estimates from one census year to the next. Reliable data on driver's licence address changes are available only from 1979 onwards in Ontario. Therefore, 1979 intercensal population estimates of Statistics Canada and estimated 1981 population were used as base. Two sets of population estimates were calculated. First, using the driver's licence address file for intra-provincial migration and second, using the income tax file for intra-provincial migration. All other components, i.e., births, deaths, interprovincial migration and international migration were kept the same for both sets of population estimates.

4.1 Errors of Closure

Two sets of population estimates for the census divisions (one using driver's licence data and the second, using income tax data for intra-provincial migration) were compared with the 1981 census population. The percent difference in the estimated population from the census population is called *error of closure*. Out of 49 census divisions, 23 have smaller errors if driver's licence data are used and 26 census divisions have smaller errors if income tax data are used to estimate intra-provincial migration.

4.2 Toronto Urban Complex

A quite interesting picture emerges in the Toronto Urban complex which includes six regional municipalities (Table 3). Driver's licence data yields a smaller error of closure for the complex as a whole and under-estimates the population for the areas outside of Metro Toronto.

The income tax file gives lower errors of closure for individual census divisions within the complex whereas, for the complex as a whole the error is larger than the driver's licence file (Table 3). Accordingly, driver licence data were used for estimating intra-provincial migration for the Toronto complex as a whole and the distribution to individual census divisions of the complex was based on revenue data.

4.3 Population Growth Rates

Percent change in the population estimates from 1979 to 1981 was calculated for the estimates derived by using the driver's licence file and the income tax file respectively. These growth rates were compared with the 1981 census growth rates.

Table 3
Errors of Closure for Toronto Urban Complex

Census Division	Errors of Closure	
	Income Tax	Driver's Licence
Durham R.M.	-0.21	-0.41
Halton R.M.	0.45	-0.51
Hamilton-Wentworth R.M.	0.10	-0.11
Peel R.M.	-0.63	-1.94
Toronto R.M.	0.01	1.26
York R.M.	-0.11	-5.70
Total Toronto Urban Complex	-0.05	-0.01

There is not much difference in the relative closeness of growth rates of the two sets to census growth rates. The number of census divisions which yield different direction of population change than the census growth rate are 3 for the driver's licence data and 10 for the income tax file (Table 4). This is one aspect where the driver's licence data appeared to yield more reliable estimates than the revenue data.

4.4 Index of Dissimilarity

Index of dissimilarity was calculated for in- and out-migration separately, as the direction of net migration was not the same for some counties for the two sets of estimates. The value of the index of dissimilarity can vary between 0 and 100. It is the half of the sum of the absolute differences between the two corresponding percent distributions and is equivalent to the sum of the positive differences or the sum of the negative differences (Shryock and Siegel 1971). The general formula is:

$$ID = \frac{1}{2} \sum |r_2 - r_1|$$

where, r_2 and r_1 are the corresponding percentages in the two distributions.

The low values of the index indicate that both files (the driver's licence and the income tax) yield quite similar estimates of intra-provincial migration for the census divisions of Ontario. However, over the four years the extent of dissimilarity increases for out-migration and improves for in-migration (Table 5).

5. CONCLUSION AND SUMMARY

This study attempts to compare the intra-provincial migration estimates derived from the driver's licence file with those derived from the income tax file. Both files provide reasonably good measures of the magnitude of intra-provincial migration for the Census Divisions of Ontario.

Although the driver's licence data appeared to provide better estimates in the direction of intra-provincial migration, the income tax data resulted in slightly more counties with smaller errors of closure and in addition yielded somewhat better results in some major areas (for example, distribution within the Toronto/Hamilton urban complex). In view of their respective strengths, the appropriate approach is to combine the use of these two data sources.

Another issue that should be noted is that the evaluation was based on three years only i.e., 1979 to 1981. A more accurate assessment on the quality of these two data files cannot be made until the availability of the 1986 census data.

To further improve the quality and the applications of the driver's licence data, the following two areas are suggested for further research:

- Verification of *FA* factor based on actual counts of unknown origin/destination through using manual coding of addresses.
- Extension of the use of the driver's licence data file as an additional source to family allowance and revenue data for inter-provincial migration estimates.

The driver's licence file tends to over-estimate migrants for Metro Toronto and under-estimate for the areas surrounding Metro Toronto. The reverse seems true for the income tax file. For this region as a whole, the driver's licence file gives better estimates for intra-provincial migration than the income tax file. The income tax file provides a better distribution of intra-provincial migrants in the counties of this region.

Table 4
Census Divisions Which Yield Different Direction of
Population Change Than Those Based on Census

Census Division	% Changed Based On	
	Income Tax	Census
Bruce	0.11	- 1.77
Grey	- 0.04	0.17
Hastings	0.12	- 1.03
Leeds and Grenville	0.21	- 0.32
Niagara	0.15	- 0.46
Northumberland	0.31	- 0.82
Oxford	0.17	- 0.09
Parry Sound	1.20	- 0.51
Stormond/Dundas/Glengarry	0.44	- 0.17
Sudbury T.D.	0.41	- 2.28
Province	1.60	1.46

	% Change Based On	
	Driver's Licence	Census
Leeds and Grenville	0.02	- 0.32
Parry Sound	0.81	- 0.51
Thunder Bay	0.42	- 0.20
Province	1.60	1.46

Table 5
Index of Dissimilarity

Year	Index of Dissimilarity	
	In-Migration	Out-Migration
1979-80	5.61	3.50
1980-81	5.54	3.82
1981-82	5.26	4.82
1982-83	4.41	4.87

REFERENCES

- ALBERTA BUREAU OF STATISTICS (1985). The development of Alberta health care records and their application to small area population estimates. A paper presented at the Meetings of the Federal-Provincial Committee on Demography, Ottawa.
- HOAG, ELIZABETH (1984). Estimating annual migration for California counties using driver's licence address change. A paper presented at the Meetings of the Population Association of America, Minneapolis, Minnesota.
- McRAE, DONALD G. (1985). The use of hydro accounts in the British Columbia based population estimates. A paper presented at the Meetings of the Federal-Provincial Committee on Demography.
- NORRIS, D., and L. STANDISH (1983). A technical report on the development of migration data from taxation records. Administrative Data Development Division, Statistics Canada.
- SHRYOCK, HENRY M., and SIEGEL, JACOB S. (1971). *The Methods and Materials of Demography*. Washington: U.S. Bureau of Census.

The Development of Alberta Health Care Records and Their Application to Small-Area Population Estimates¹

**F. AHMAD, R. CHOW, O. DEVRIES,
A. HASHMI, and M. MARCOGLIESE²**

ABSTRACT

This paper examines the use of administrative files from Alberta's Health Care Insurance Plans combined with Vital Statistics data as inputs for estimating population. Results, which are presented and compared with Census data, indicate that Health Care data can be used to produce accurate population estimates at the provincial level and for smaller areas such as census divisions and municipalities.

KEY WORDS: Administrative files; Component method; Small areas; Residual net migration.

1. BACKGROUND

During the mid to late 1970's, the Province of Alberta experienced rapid economic growth led by activity in the oil and gas industry, which generated high population growth. Governments, in order to effectively provide goods and services for the influx of people into various regions, required timely data on where and by how much population was growing. With the need for up-to-date population data, it was felt that the federal quinquennial census was not sufficiently frequent nor current (census data are released about twelve to eighteen months after the reference year). Consequently, provincial agencies, and in particular, the Alberta Bureau of Statistics, began investigating alternative sources of timely population data.

After examining a number of potential sources, the Bureau began assessing administrative health care insurance data from the Alberta Health Care Insurance Plan (AHCIP) files to develop population statistics. The remainder of this paper highlights work undertaken by the Bureau to develop the AHCIP records and to use the data in estimating small-area population.

2. DEVELOPMENT OF AHCIP RECORDS INTO HEALTH CARE COUNTS

This section describes briefly the nature of the AHCIP records and evaluates the counts developed.

2.1 Developing Health Care Counts Data

The Bureau receives selected registration records via computer tape, on a quarterly basis, from the AHCIP registration-billing system. (The tape contains only a partial listing, in particular, all names, identifiers, etc. have been stripped such that the confidentiality of all individuals is strictly preserved.) The file contains information such as addresses, postal codes, registration and cancellation dates, age and sex for every registrant. (A detailed description of the record layout is available upon request.)

¹ Abridged version of the paper presented at the Federal-Provincial Committee on Demography meeting held on November 26-27, 1985, Ottawa, Canada.

² F. Ahmad, R. Chow, O. DeVries, A. Hashmi and M. Marcogliese, Alberta Bureau of Statistics, Alberta Treasury, Sir Frederik W. Haultain Building, 9811-109th Street, Edmonton, Alberta, Canada T5K 0C8.

The reporting unit of the AHCIP file is the registration. Each registration may contain up to twenty-five individuals; one registrant (usually the person who pays the premiums) and up to twenty-four dependents. There are currently about 1.7 million active registrations accounting for roughly 2.6 million individuals. In addition, the file is historical and includes all individuals ever covered under AHCIP since its inception in 1969.

The file is processed through four phases.

- a) Edit-notes and/or corrects errors according to edit check criteria.
- b) Purge-uses the edited raw data file and selects active individuals.
- c) Consolidation-matches postal codes between the purged file and the Bureau's Postal Code Translator File (PCTF) and attaches the geographic reference information to the AHCIP records.
- d) Aggregation-takes the consolidated file and aggregates males and females by single years of age for each postal code. This reduces the number of records/individuals from approximately 2.6 million to fewer than 120,000 and significantly reduces the subsequent systems processing costs.

The aggregated file is used for the production of age and sex counts by any geographic area definable through the 60,000 PCTF Alberta codes.

2.2 Evaluation of the Counts Data

To evaluate the health care counts data, Census of Canada population figures for 1976 and 1981 were used for comparison. The 1981 AHCIP records were considered to be more accurate than the 1976 file, therefore, the evaluation relied more heavily upon the 1981 census comparisons. Also used as a second basis of comparison were municipal censuses data, even though these data generally were not considered to be as reliable as Canada Census figures. The municipal censuses, however, provided insight into the magnitude of the variations as well as the relative distributions of age, sex and trends (growth or decline) over time. An additional source of comparison was intercensal population estimates prepared by the Bureau and by Statistics Canada.

Basic findings:

- a) On a provincial basis, AHCIP counts overestimate both Canada Census and total municipal censuses figures by about 3.5% to 4.5%. Age and sex distributions are more accurate and the correlation coefficients indicate consistency of trends (over/under estimates) over time.
- b) At the census division (CD) level, AHCIP counts varied from Canada Census figures from -2.6% to 9.7% (see Table 1). Comparisons with intercensal population estimates indicated a similar variance. As with the provincial level data, age and sex distributions and the trend consistency proved highly reliable.
- c) At the census consolidated subdivision (CCSD) level, for fifty of the seventy-one CCSDs, health care data were within $\pm 10\%$ of the Census counts. The largest discrepancy was -56.5% (Municipal District 135).

Most problem areas had major urban centres located close to the county, municipal district and improvement district boundaries. No specific anomalies were found when testing the age and sex distributions, although relationships were not as strong as with the province and the census division levels.

- d) At the census subdivision (CSD) level, preliminary figures showed discrepancies between the AHCIP counts and 1981 Census data ranged from -100% to +95%.

Consequently, the twenty-eight largest areas of over 5,000 in population were used at the CSD level. The six largest CSDs (Edmonton, Calgary, Lethbridge, Medicine Hat, Red Deer and St. Albert) displayed overcounts ranging from 3% to 9%. Eight other CSDs differed up to $\pm 20\%$, while sixteen showed somewhat greater than $\pm 20\%$ variation. Again, no specific age and sex distribution anomalies were detected, although discrepancies were greater than those at more aggregated levels. As well, twenty-seven of the twenty-eight CSDs indicated high trend consistency.

As the geographic area decreases in size, AHCIP counts become less reliable; age and sex distributions, although less accurate, still remain strong; and trend consistency (counts over time) remain highly correlated with a few notable exceptions. The limitations of AHCIP counts as population indicators primarily can be attributed to one of two main sources: a) the AHCIP administrative procedures/inaccuracies; or b) use of postal codes.

a) *AHCIP Administrative Procedures:*

- 1) As an insurance programme, a chief concern is to supply coverage. Therefore, efforts are directed to getting people onto the system to ensure universal coverage with less effort placed on getting individuals off the system. This has resulted in more people being registered than are actually in the province.

Table 1
Comparisons of Alberta Health Care Counts and Canada Census Data
for Alberta Census Divisions

Census Division	Year							
	1976				1981			
	Census Count	AHCIP Count	Percent Difference Count	Actual Difference Count	Census Count	AHCIP count	Percent Difference Count	Actual Difference Count
1	46,990	45,789	-2.56	-1,201	55,375	55,748	0.67	373
2	96,995	97,229	0.24	234	110,477	111,567	0.99	1,090
3	32,898	33,884	3.00	986	35,652	36,463	2.27	811
4	12,130	12,101	-0.24	-29	12,119	12,038	-0.67	-81
5	35,424	35,656	0.65	232	38,382	38,457	0.20	75
6	524,554	538,432	2.65	13,878	668,682	699,999	4.68	31,317
7	37,866	38,235	0.97	369	40,071	40,359	0.72	288
8	95,384	95,063	-0.34	-321	123,642	124,666	0.83	1,024
9	19,903	21,832	9.69	1,929	21,670	23,338	7.70	1,668
10	67,171	67,168	0.00	-3	78,417	78,532	0.15	115
11	632,909	646,799	2.19	13,890	762,041	796,884	4.57	34,843
12	63,129	62,011	-1.77	-1,118	84,221	86,183	2.33	1,962
13	46,305	47,258	2.06	953	53,701	54,282	1.08	581
14	19,386	21,039	8.53	1,653	24,635	25,991	5.50	1,356
15	106,993	111,678	4.38	4,685	128,639	134,451	4.52	5,812
Unknown ^a		48,462				19,279		
Alberta	1,838,037	1,922,636	4.60	84,599	2,237,724	2,338,237	4.49	100,513

^a Unknown, represent counts without address identifiers.

Source: Statistics Canada 1976 and 1981 Censuses; Alberta Health Care Insurance Plan data, prepared by Alberta Bureau of Statistics, Alberta Treasury.

- 2) Mailing addresses are used rather than residential addresses, which has created difficulties in assigning geographic locations. Discrepancies occur in areas where significant rural populations surround an urban centre and the rural populace pick up their mail in the urban centre. Consequently, most urban areas are overcounted while rural areas are undercounted.
- 3) Incomplete and inaccurate data, especially related to postal codes, make it difficult to produce small-area statistics due to undercounting.
- 4) Time lags in reporting and recording of the data influence counts. Generally speaking, it takes three to six months to get an individual onto the system (birth, in-migrant) but it requires usually much longer to be removed from the active system (death, out-migrant). The lags, however, are difficult to follow and differ substantially depending on the circumstances.

b) *Postal Codes:*

- 1) Postal codes define delivery service areas (where a person gets his mail), not necessarily a residence. This factor limits the accuracy of assigning AHCIP registrations to appropriate geographic areas. In particular, it creates urban-rural split problems, as discussed.
- 2) A six-digit postal code, by itself, is not always enough to determine the service delivery area. A rural route, suburban service, or box number may be required to further specify a more exact location.
- 3) Postal codes have been insufficient, especially in rural areas, to aggregate to appropriate levels. For example, there are approximately 363 census subdivisions in Alberta, but the Bureau's PCTF can derive only 324 of these.

The problems outlined above have precluded the release of AHCIP counts as approximations of actual population. Although the counts were quite good in some areas, in others, they were poor or inconsistent. With the strong relationships between health care, age and sex distributions and those of Canada Census, as well as the consistency of trends over time, the counts have been used in conjunction with the Bureau's population estimation methodology (as discussed in the next section).

3. APPLICATION OF HEALTH CARE COUNTS TO SMALL-AREA POPULATION ESTIMATES

The Bureau has produced intercensal population estimates for Alberta and provincial census divisions for nearly a decade. During this period, various methodologies and data sources have been examined and used to improve the quality of these estimates. To date, significant success has been achieved with the component method using health care counts as input data. These data have been used to derive the age and sex structure of the Alberta population at the provincial and census division level and to produce provincial and census division population estimates. Also, recently, the data have been used to test the applicability in preparing census subdivision population estimates.

3.1 Estimation Methodology

The estimation methodology employed by the Bureau to produce subprovincial population estimates is comprised of two parts. Part one presents the method of estimating migrant population. Part two outlines the method used to develop population estimates.

a) *Estimating Migrant Population Using Health Care Counts*

The Bureau developed data from three administrative files: counts from AHCIP records; births from data supplied by Alberta Vital Statistics; and deaths, also supplied by Alberta Vital Statistics. These sources were used to calculate net migration. Basically for any small area, the growth of health care counts is obtained from the differences in counts between time t and time $t-1$. This residual less the area's natural increase (births minus deaths) calculates the inflow (or outflow) of individuals, i.e., net migration. This procedure is mathematically expressed as:

$$HMIG = [(HC_t - HC_{t-1}) - (B - D)]$$

Where:

$HMIG$ = health care net migration counts between time t and $t-1$

HC_t = total health care counts at time t

HC_{t-1} = total health care counts at time $t-1$

B = total births during time interval t to $t-1$

D = total deaths during time interval t to $t-1$.

This health care migrant population estimate, however, is subject to the same over and under counting difficulties discussed in Section 2. As a result, although this approach would prepare estimates for small areas at the provincial level, these estimates would be less reliable than the provincial migration estimates currently derived using interprovincial flows to family allowance recipients. (The family allowance files are also used by Statistics Canada, which ensures provincial estimates generally are consistent with those produced at the federal level.)

To further improve the small-area migration estimates and to ensure consistency with estimates at the provincial level, should the small areas be aggregated to a provincial total, an adjustment using a ratio distribution was encompassed. With this approach, the ratio of net migration from health care counts for an area over the net migration from health care counts for the province is multiplied by the provincial net migration calculated in connection with the Bureau's quarterly population estimates. Mathematically, the equation is:

$$AMIG_i = \frac{HMIG_i}{HMIG_a} \times PMIG$$

Where:

$AMIG_i$ = adjusted net migration in area i

$HMIG_i$ = health care net migration of counts for area i

$HMIG_a$ = health care net migration of counts for Alberta

$PMIG$ = estimated provincial net migration from Alberta's quarterly population estimates.

This adjusted migration estimate (AMIG) is then used as input into estimating population.

b) *Estimation of Population For Small Areas*

The adjusted estimated net migration (AMIG) for each area is used in an equation using the components of population growth (births, deaths and migration):

$$P_{i_t} = P_{i_{t-1}} + (B_i - D_i) + AMIG_i$$

Where:

P_{i_t} = estimated population in area i at time t

$P_{i_{t-1}}$ = population in area i at time $t - 1$

3.2 Evaluation of Small-Area Estimates

Using the above approach, the Bureau has developed population estimates for Alberta's fifteen census divisions and twenty-eight municipalities with populations over 5000. The results, so far, have been promising.

The results of a comparison between 1981 census data and estimates for 1981 prepared with 1976 census figures as a base population using the above described methodology, at the census division level, are presented in Table 2. For thirteen of the fifteen divisions the estimates were within $\pm 2.0\%$ variation compared to the 1981 census. Only the two smallest CDs (9 and 14) showed a five-year deviation greater than 2.0%. The average absolute deviations (i.e., average annual deviations) were no greater than 0.5% for all census divisions.

The twenty-eight population estimates for municipalities were compared to the 1981 census counts, as well as available data from municipal censuses conducted from 1982 to 1984 (Tables 3 and 4). Federal census comparisons showed nineteen estimates of the twenty-eight municipalities had an average absolute deviation of less than $\pm 1.0\%$. Only six municipalities had annual differences greater than 2.0%. Comparisons with municipal censuses conducted between 1982 and 1984, yielded twenty-two instances of deviations within $\pm 1.0\%$, fourteen ranging between $\pm 1.0\%$ and $\pm 3.0\%$, while nine had deviations greater than $\pm 3.0\%$.

In general, the estimation results have been satisfactory and encouraging. The development of AHCIP registrant counts and the component approach employed to estimate population have improved the accuracy of the population estimates produced and opened up possibilities for deriving estimates for user-defined small geographic areas. The Bureau will continue to investigate ways to improve the AHCIP counts (some of which are related to new administrative procedures being incorporated for the AHCIP). Also, the population estimation methodology will be further refined as new data techniques become available.

3.3 Summary of Advantages and Disadvantages of Using AHCIP

Using health care counts in deriving small-area population estimates has a number of advantages and disadvantages.

Table 2
Comparisons of Canada Census Counts and Alberta Bureau of Statistics
Population Estimates for Alberta Census Divisions

Census Division	Census 1976	Bureau Estimates ^a			
		Natural Increase ^b 1976-81	Net Migration 1976-81	Growth 1976-81	Population 1981
1	47,000	2,730	6,080	8,810	55,810
2	96,980	6,120	7,190	13,310	110,290
3	32,870	2,310	100	2,410	35,280
4	12,140	490	- 520	- 30	12,110
5	35,460	1,820	790	2,610	38,070
6	524,570	33,860	107,540	141,400	665,970
7	37,820	2,010	- 10	2,000	39,820
8	95,400	6,140	20,860	27,000	122,400
9	19,850	1,040	200	1,240	21,090
10	67,230	1,650	8,550	10,200	77,430
11	632,830	43,880	90,880	134,760	767,590
12	63,130	6,470	16,130	22,600	85,730
13	46,300	2,040	4,320	6,360	52,660
14	19,450	2,200	2,430	4,630	24,080
15	107,010	10,260	10,040	20,300	127,310
Alberta	1,838,040	123,020	274,580	397,600	2,235,640

Census Division	Census 1981	Difference		Average Absolute Deviation
		Number	%	
1	55,360	450	0.81	0.16
2	110,470	- 180	- 0.16	0.03
3	35,640	- 360	- 1.01	0.20
4	12,120	- 10	- 0.08	0.02
5	38,430	- 360	- 0.94	0.19
6	668,680	- 2,710	- 0.41	0.08
7	40,030	- 210	- 0.52	0.10
8	123,690	- 1,290	- 1.04	0.21
9	21,630	- 540	- 2.50	0.50
10	78,390	- 960	- 1.22	0.24
11	762,080	5,510	0.72	0.14
12	84,220	1,510	1.79	0.36
13	53,690	- 1,030	- 1.92	0.38
14	24,650	- 570	- 2.31	0.46
15	128,640	- 1,330	- 1.03	0.21
Alberta	2,237,720	- 2,080	- 0.09	0.02

^a Data are experimental.
^b Natural increase refers to the number of births minus the number of deaths.
Note: Components may not add to total due to rounding.
Source: Statistics Canada 1976 and 1981 Censuses; Alberta Bureau of Statistics Estimates.

Table 3

Comparisons of the Canada Census Counts and Alberta Bureau of Statistics
Population Estimates for Selected Alberta Municipalities

Municipality	Census 1976	Bureau Estimates ^a			Census 1981	Difference %	Average Absolute Deviation
		Natural Increase ^b 1976-1981	Net Migration 1976-1981	Popu- lation 1981			
Airdrie	1,410	580	5,090	7,070	8,410	-15.9	3.2
Brooks	6,340	730	2,370	9,440	9,420	0.2	0.0
Calgary	469,920	30,310	93,760	593,990	592,740	0.2	0.0
Camrose	10,100	150	2,570	12,830	12,570	2.1	0.4
Crowsnest Pass	5,250	40	-410	4,880	7,310	-33.2	6.6
Drayton Valley	4,300	530	1,760	6,590	5,040	30.8	6.2
Drumheller	6,150	20	220	6,390	6,510	-1.8	0.4
Edmonton	461,360	27,900	51,240	540,510	532,250	1.6	0.3
Edson	4,040	510	2,490	7,040	5,840	20.5	4.1
Fort McMurray	15,420	2,900	14,140	32,460	31,000	4.7	0.9
Fort Saskatchewan	8,300	800	2,660	11,760	12,170	-3.4	0.7
Grande Prairie	17,630	1,970	6,300	25,900	24,260	6.8	1.4
Hinton	6,730	760	-820	6,670	8,340	-20.0	4.0
Innisfail	2,900	230	1,930	5,060	5,250	-3.6	0.7
Lacombe	3,890	150	1,210	5,240	5,590	-6.3	1.3
Leduc	8,580	920	3,430	12,930	12,470	3.7	0.7
Lethbridge	46,750	2,070	4,400	53,220	54,070	-1.6	0.3
Medicine Hat	32,810	1,770	6,010	40,590	40,380	0.5	0.1
Peace River	4,840	580	970	6,390	5,910	8.1	1.6
Ponoka	4,640	-10	530	5,160	5,220	-1.1	0.2
Red Deer	32,180	2,300	11,790	46,270	46,390	-0.3	0.1
Spruce Grove	6,910	1,110	4,710	12,730	10,330	23.2	4.6
St. Albert	24,130	2,360	6,670	33,160	32,000	3.6	0.7
Stettler	4,180	500	580	5,270	5,140	2.5	0.5
Taber	5,300	320	410	6,020	5,990	0.5	0.1
Vegreville	4,160	80	860	5,090	5,250	-3.0	0.6
Wetaskiwin	6,750	300	2,440	9,490	9,600	-1.1	0.2
Whitehorse	3,880	600	1,150	5,630	5,590	0.7	0.1
Alberta	1,838,040	123,020	274,580	2,235,630	2,237,720	-0.1	0.0

^a Data are experimental.

^b Natural increase refers to the number of births minus the number of deaths.

Note: Components may not add to total due to rounding.

Source: Statistics Canada 1976 and 1981 Censuses; Alberta Bureau of Statistics Estimates.

Table 4

Comparisons of Alberta Municipal Censuses and Alberta Bureau of Statistics
Population Estimates for Selected Municipalities

Municipality	1982			1983			1984		
	Bureau Esti- mate ^a	Muni- cipal Census	Devia- tion %	Bureau Esti- mate ^a	Muni- cipal Census	Devia- tion %	Bureau Esti- mate ^a	Muni- cipal Census	Devia- tion %
Airdrie	9,450	9,980	-5.3	9,830	10,430	-5.8	10,080	--	--
Brooks	9,640	--	--	9,790	--	--	9,510	--	--
Calgary	614,930	623,130	-1.3	622,510	620,690	0.3	615,140	619,810	-0.8
Camrose	12,880	12,810	0.6	12,970	--	--	13,070	12,750	2.5
Crowsnest Pass	7,490	7,580	-1.1	7,530	--	--	7,350	--	--
Drayton Valley	5,120	4,870	5.2	5,200	--	--	5,310	4,920	7.9
Drumheller	6,660	--	--	6,700	6,670	0.4	6,620	--	--
Edmonton ^b	550,930	551,310	-0.1	557,400	560,090	-0.5	551,140	--	--
Edson ^b	6,110	6,290	-2.9	6,220	--	--	6,080	7,110	-14.5
Fort McMurray	32,930	33,580	-1.9	33,600	34,490	-2.6	35,150	35,350	-0.6
Fort Saskatchewan	12,530	12,460	0.6	12,650	12,470	1.4	12,620	--	--
Grande Prairie	24,650	--	--	24,910	24,080	3.5	25,370	24,410	3.9
Hinton	8,820	8,820	0.0	8,980	8,830	1.8	8,950	8,900	0.6
Innisfail	5,420	5,440	-0.4	5,460	--	--	5,440	5,440	0.0
Lacombe	5,810	5,720	1.5	5,850	5,850	5,950	-1.8	5,850	--
Leduc	12,880	--	--	13,010	--	--	13,290	--	--
Lethbridge ^b	55,440	56,500	-1.9	55,900	58,090	-3.8	57,500	--	--
Medicine Hat ^b	41,070	--	--	41,440	42,270	0.7	41,540	--	--
Peace River	6,080	--	--	6,150	--	--	6,250	--	--
Ponoka	5,310	--	--	5,310	--	--	5,280	--	--
Red Deer	48,450	48,560	-0.2	49,230	50,260	-2.0	50,860	51,070	-0.4
Spruce Grove	11,080	10,780	2.7	11,410	11,310	0.9	11,550	11,570	-0.1
St. Albert	33,170	32,980	0.6	33,740	35,030	-3.7	34,840	35,530	-1.9
Stettler	5,180	--	--	5,220	--	--	5,300	--	--
Taber	6,140	--	--	6,210	--	--	6,360	6,380	-0.4
Vegreville	5,280	5,250	0.6	5,290	--	--	5,390	--	--
Wetaskiwin	9,880	9,900	-0.2	9,990	10,020	-0.3	10,080	--	--
Whitecourt	5,710	--	--	5,840	--	--	5,710	--	--

^a Data are experimental.

^b Annexation took place between 1982 and 1984.

Note: "--" indicates that a municipal census is not available.

Source: Alberta Municipal Affairs, 1982-1984 Municipal Censuses; Alberta Bureau of Statistics Estimates.

Advantages:

- a) AHCIP registration data provides universal coverage of all individuals in Alberta;
- b) Registration lag appears to be random and does not adversely affect distributions or trends of the counts;
- c) Data are available on a timely/frequent basis; and
- d) The file contains some socio-economic information on registrants and dependents (e.g., age, sex and marital status) to enable the production of more than basic population estimates.

Disadvantages:

- a) Residency based on postal codes can lead to some inaccuracies;
- b) AHCIP registrants can leave the system, for example, death and out-migration, without notifying AHMC resulting in overcounts; and
- c) Administrative procedures may cause discrepancies/inaccuracies in the number of Alberta Health Care registrants.

4. CONCLUSION

Our experience with health care development has been very positive. The greatest potential is the use of the counts in a component model to produce estimates for small areas as well as the excellent age-sex distribution ratios and trend consistency. Costs of development of the demographic reporting systems were not considered excessive in light of these benefits. For other provincial agencies contemplating the development of provincial health care files, the Bureau would certainly be willing to discuss its experiences in more detail and make available additional information, such as record layouts and system processing costs.

The Use of Hydro Accounts in the British Columbia Regression Based Population Estimation Model¹

DONALD G. McRAE²

ABSTRACT

The accuracy of small area population estimates derived from a regression based model is heavily dependent on the ability of the indicator data selected to accurately reflect population change. Hence, prior knowledge as to the characteristics of the administrative data used as potential population indicators in a regression model is important. This report summarizes the strengths and weaknesses associated with the use of residential hydro accounts in the British Columbia regression based population estimation model.

KEY WORDS: Small area population estimates; Regression method; Difference-correlation method; Population indicators; Hydro accounts; Family allowance recipients.

1. INTRODUCTION

The Central Statistics Bureau produces post-censal population estimates for a variety of geographic units within the Province of British Columbia including municipalites, local health areas, census divisions and RCMP regions among others. Current population estimates are produced for these sub-provincial areas by means of a regression approach, specifically the Difference-Correlation Method (DCM).

A detailed description of this methodology is given in earlier papers (Central Statistics Bureau 1982, McRae 1985). The data used as indicators of population are the number of family allowance recipients (F), and/or the number of residential hydro accounts (H). The characteristics of this second data source, residential hydro accounts, relative to the British Columbia model will be examined over the remainder of this paper.

2. DATA SOURCES

Residential hydro accounts data within British Columbia are obtained from nine different organizations. These are:

Organization	% of Total Hydro Accounts (1985)
(1) British Columbia Hydro	90.9
(2) West Kootenay Power and Light	4.7
(3) Princeton Light and Power Co.	0.2
(4) City of Kelowna	0.8
(5) City of Penticton	0.8
(6) District of Summerland	0.3
(7) City of New Westminster	1.7
(8) City of Grand Forks	0.1
(9) City of Nelson	0.6

¹ Presented at the meeting of the Federal-Provincial Committee on Demography, Ottawa, November 28-29, 1985.

² D.G. McRae, Central Statistics Bureau, Ministry of Industry and Small Business Development, Government of British Columbia, 2nd Floor, 1405 Douglas Street, Victoria, British Columbia, Canada V8W 3C1.
The views expressed in this paper are those of the author and do not necessarily represent the views of the Government of British Columbia.

The major suppliers of residential electrical power are British Columbia Hydro and West Kootenay Power and Light. The other organizations purchase power from the two major suppliers, and retail this electricity to their own customers (usually the residents of the municipality).

3. DATA FORMAT

Of the nine sources of residential hydro accounts data, only that provided by British Columbia Hydro is in machine readable form. The other eight organizations provide the data totalled by municipality (urban), along with a total of any rural (non-municipal) customers. The reference date for all data is the May 31 billing file, and in most cases the data can be obtained within 2 to 3 weeks of the billing date.

Data provided by British Columbia Hydro is in two formats. The first shows the number of residential meters as of May 31 by Capital District Code. A Capital District Code, of which there are approximately 248 in the Province, is an administrative unit used by British Columbia Hydro and corresponds to a municipality where municipalities exist. By agreement, both the major power suppliers in the Province pay each municipality a certain percentage of the annual revenue collected from the residential customers in that municipality in lieu of property taxes. As a result, power companies such as British Columbia Hydro, design their accounting systems to correspond to customers within municipal boundaries. In addition, British Columbia Hydro attempts to maintain a close correspondence between Capital District boundaries and school district boundaries.

The second format provides for each of the one million plus residential meters the postal code of billing address. This second data file allows the easy translation of hydro meters to geographic units other than municipalities and school districts via the postal code.

4. STRENGTHS OF THE HYDRO DATA IN A REGRESSION MODEL

Empirical tests of the two different data sources, hydro meters (H) and family allowance accounts (F), were conducted by producing 1981 population estimates with each separately and together. The regression coefficients used were derived from the pooled 1971/76 and 1976/81 periods, and the base year was 1976. The results were compared with the 1981 Census, and the Average Absolute Percent Errors (AAPE) were calculated. The results are given in Tables 1 and 2.

As can be seen in Table 1, population estimates based on hydro data produce, on average, lower percentage errors than the family allowance based estimates. Closer examination of Table 1 reveals that the improvement in estimation accuracy lies almost entirely with the estimates for areas with population less than 4000. This observation is reinforced in Table 2 where it is shown that, statistically speaking, there is a significant difference in the estimation accuracy between the hydro and family allowance based estimates for areas less than 4000 population.

The marginal effect of adding another population indicator to the Difference-Correlation Method can be judged by examining the change in estimation accuracy with and without the additional indicator. It would appear from Tables 1 and 2 that the inclusion of hydro data statistically improves the estimation accuracy in both large and small areas. Family allowance data, on the other hand, improves the accuracy for larger areas but reduces the estimation accuracy for smaller areas, with no statistically significant effect overall.

Table 1
Comparison of Estimation Errors Among Data Sources
for British Columbia Municipalities - 1981

Data Source	AAPE Overall	n	Population			
			≥ 4000 AAPE	n	< 4000 AAPE	n
DCM/H/F	5.53	158	2.99	88	8.72	70
DCM/H	5.16	158	4.04	88	6.58	70
DCM/F	10.46	167	4.57	92	17.69	75

$$AAPE = \left[\sum_{i=1}^n \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right| \right] \div n \times 100$$

where:

- Y_i = census population for region i
- \hat{Y}_i = estimated population for region i
- n = number of areas estimated.

Table 2
Test for Statistically Significant Differences Between the Average
Absolute Percent Errors for Selected Data Sources - 1981

Data Source	Overall	95% Confidence Interval for the Average Difference in Absolute Percent Errors	
		Population	
		≥ 4000	< 4000
DCM/H/F - DCM/H	.37 ± .86	-1.05 ± .56 ^a	2.14 ± 1.76 ^a
DCM/H/F - DCM/F ^b	-4.86 ± 1.55 ^a	-1.57 ± .85 ^a	-9.00 ± 3.11 ^a

^a Statistically significant differences at the 5% level utilizing a two tailed T-test, paired samples and assuming normally distributed means.

^b In order to pair the samples only 158 of the possible 167 family allowance base estimates were used. The number of observations were: overall, 158; greater than or equal to 4000, 88; less than 4000, 70.

5. WEAKNESSES OF HYDRO DATA IN A REGRESSION MODEL

One problem encountered when using hydro data in a regression model for population estimates is that of vacant dwellings, or more accurately, significant differences in the rate of vacant dwellings between the base and estimating years. This weakness of the data was demonstrated by the 1981 evaluation of the communities in the Peace River-Liard region of British Columbia (McRae 1982). As a result of the North-East Coal project, the communities of the Peace River-Liard Census Division in 1981 were experiencing a building boom as developers constructed dwellings in anticipation of a population influx. Each dwelling, occupied or not (or even under construction) would require a meter, which may have had low usage, but was still active and hence counted. As a result, the change in share of meters from 1976 to 1981 was overstated relative to population, producing overestimates of the 1981 population for many of Peace-River Liard communities.

Another weakness of the hydro data is the potential for a change-over of multiple dwelling units from single to multiple meters. This may occur when an older apartment building, for example, serviced by a single meter is remodelled or replaced with individually metered units. This problem would produce an overestimate of population in a regression model if it were to occur sometime between the base and estimating years.

Finally, some problems will result if the hydro data is used for areas that have a changing nature, or in other words, a changing relationship between population and residential meters. One example of this in B.C. is the resort community of Whistler. Fifteen years ago this municipality was largely a collection of winter cabins on a ski hill. However, over the last decade this area has been shifting to a year round residence basis. Consequently, the number of persons per hydro meter, which was originally very low relative to the B.C. average, is moving toward the norm. Like the vacancy problem, the use of hydro data to estimate the population for such a community would likely result in above average errors.

The solution to all three of the problems mentioned above is to remove accounts that have a low monthly or bi-monthly usage, and hence are assumed to be vacant. The feasibility of this procedure is currently being examined by the Central Statistics Bureau in relation to the data obtained from B.C. Hydro. If possible, we hope to have the improved data set available for calibration against the 1986 Census. Currently, as a partial solution hydro data for areas that had in 1981 a low or high ratio of persons per meter relative to the provincial norm (i.e. less than 2 or greater than 5) are not used.

A final potential weakness of the hydro data is the reliance on external and different organizations for the information. In the past, this situation generally has not proven to be a problem. However, there have been some rare cases that have called into question the quality of the meter data collected in the field. Such a case may be a boundary change of a municipality not being reflected in the meter data, or the addition of some types of non-residential accounts (such as lamp standards) to the data. As a result, careful monitoring of the data is important.

6. CONCLUSIONS

The following strengths and weaknesses are associated with the use of hydro data in the British Columbia regression based population estimation model.

Strengths:

- (a) The hydro data, when used in a regression model, produces a lower average absolute percent error than family allowance data for small areas.
- (b) The data is obtained from each supplier in a format that is already aggregated to municipalities. The major advantage of this is that changes in municipal boundaries, which occur regularly, are reflected in the data with no additional work on the part of the Bureau.
- (c) The majority of the data can be obtained in machine readable form along with the postal code. This allows the easy translation of the data to geographic regions other than municipalities when sorted by the Bureau's postal code Translation Master File.
- (d) The data can be obtained free of charge from each of the suppliers within a relatively short time period (2 to 3 weeks).

Weaknesses:

- (a) Differential vacancy rates between the base and estimating years will bias the estimates.
- (b) Dwelling units (such as apartments) that change from a single to multiple meter sometime between the base and estimating years will bias the estimates upwards.

- (c) Areas with a changing nature, such as from a seasonal to "stable" population, will introduce bias into the estimates.
- (d) The data is obtained from external and different organizations. This potentially could cause problems in terms of data quality and comparability, as well as producing a situation in which the priorities of the Bureau's population estimates program are subservient to the administrative needs of an external organization.

REFERENCES

- CENTRAL STATISTICS BUREAU (1982). British Columbia municipal population estimation methodology. Unpublished Report, Central Statistics Bureau, Ministry of Industry and Small Business Development, Government of British Columbia.
- McRAE, D. (1985). A regression approach to small area population estimation. Paper submitted to the International Symposium on Small Area Statistics, Ottawa, Canada.
- McRAE, D. (1982). British Columbia small area estimation model - 1981 municipal and census division evaluation. Unpublished Report, Central Statistics Bureau, Ministry of Industry and Small Business Development, Government of British Columbia.

Estimating the Age/Sex Distribution of Small Area Populations¹

DAVID S. O'NEIL and CHRIS D. McINTOSH²

ABSTRACT

This paper describes a method of producing current age/sex specific population estimates for small areas utilizing as inputs total population estimates, birth and death data and estimates of historical residual net migration. An evaluation based on the 1981 Census counts for census divisions and school districts in British Columbia is presented.

KEY WORDS: Age/sex population estimates; Small area; Residual net migration.

1. INTRODUCTION

The Central Statistics Bureau currently produces post-censal population estimates for a variety of sub-provincial areas using a regression approach (Central Statistics Bureau 1982). In addition to estimates of the total population by small area, age/sex specific estimates are also produced.

This paper outlines the method by which age/sex specific population estimates are derived for subprovincial areas of British Columbia, given an estimate of the total population.

2. OVERVIEW

The methodology used to derive the small area populations by sex and single years of age is divided into two parts.

The first part consists of examining historical residual net migration data compiled from censuses to derive a number of migration distributions by sex and single year of age for each small area (Shryock and Siegal 1980).

The second part of the methodology consists of aging the base population for each sex and adding births and subtracting deaths to yield a new population distribution for each area. This is referred to as the "natural base" population. The difference between the estimated total population by sex and the natural base population yields a residual term, which is equal to net migration by sex if the population and vital events for the two periods are exact. This small area sex specific residual term is distributed by single years of age according to a historical distribution, then added to the natural base population giving an age/sex specific population estimate for the area in the next time period.

Due to the timeliness of the input data, estimates of the total populations can be produced four months after the reference date of June 1, and the age/sex breakdowns one to two months later.

¹ Abridged version of the paper presented at the meeting of the Federal-Provincial Committee on Demography, Ottawa, November 28-29, 1985.

² D.S. O'Neil, SRL Sociometrics Resources Ltd., and C.D. McIntosh, Intersoft Resources Ltd., Central Statistics Bureau, Ministry of Industry and Small Business Department, Government of British Columbia, 2nd Floor, 1405 Douglas Street, Victoria, British Columbia, Canada V8W 3C1.
The views expressed in this paper are those of the authors and do not necessarily represent the views of the Government of British Columbia.

3. HISTORICAL NET MIGRATION DISTRIBUTIONS

Age/sex specific residual estimates of net migration were compiled for the census periods 1961/66, 1966/71 and 1971/76 for each of the 74 British Columbia school districts. These are referred to as the Historical Small Area Distributions.

Examination of these net migration distributions by small area showed them to be extremely unstable over time. In order to minimize the effects of this instability, a number of steps were taken.

First, migration distributions by small area were separated according to whether they occurred during a time of positive or negative total net migration. It was found that residual migration age distributions for many areas differed depending on whether net migration was positive or negative.

A further step taken to reduce the effects of unstable migration distributions was to group small areas of similar proportional migration distributions together, then calculate the positive and negative net migration distributions for each group of areas. These were called the Historical Grouped Distributions. Cluster analysis (using the SPSS/PC procedure) across selected age groups was used to group the historical small area migration distributions. Examination of cluster memberships from different periods resulted in the placing of the majority of areas into three clusters, while eight areas were maintained as unique independent clusters. Once areas had been arranged into groups, positive and negative migration distributions were calculated from the most recent periods of positive or negative net migration.

4. SMALL AREA POPULATION ESTIMATES BY SEX AND SINGLE YEAR OF AGE

As noted in Section 3, some areas showed considerable time-series variation in the residually calculated net migration distributions. This was likely the result of two factors. First, many of the areas under study possess small resource based economies subject to wide fluctuations, with consequent swings in migration levels. Second, a certain amount of instability is introduced when calculating a percentile distribution for a concept such as net migration, which may have either positive, negative, or zero values.

In order to guard against adopting a historical net migration distribution that may not be a representative distribution for the estimating year, five different historical sex-specific distributions were calculated, then distributed by single year of age. A description of these five different net migration distributions is given below.

- 1) The Historical Small Area Distribution for each small area having the same sign as the net migration to that small area was the first migration distribution.
- 2) The Historical Group Distribution for the group the small area belongs to, having the same sign as the net migration to that small area, was the second migration distribution.
- 3) The third migration distribution was calculated by separately totaling the migration from the most recent time period for all small areas with a positive and negative net migration, then calculating the age distributions.
- 4) The fourth distribution was the distribution of the natural base population for each small area.
- 5) The fifth and final distribution was the age distribution of migrants to British Columbia as a whole. For all the years under consideration, migration to B.C. has been positive, hence this is a positive distribution. Nevertheless, it was used as the fifth distribution regardless of whether the migration to a small area was positive or negative.

In some cases it was not possible to calculate all five distributions. This was the case if a small area never had a negative net migration in the past, but one is indicated for the estimating year under consideration. In situations such as this only distributions that can be calculated were used to distribute the small area net migration.

Empirical testing based on the 1981 Census indicated that of the five net migration distributions described above, number 1 (the Historical Small Area Distribution) produced the lowest average absolute percent error over all school districts and age groups, followed by number 2 (Historical Grouped Distribution), then number 3, etc. However, despite the fact that distribution number 1 produced the lowest error on average, it did not produce the lowest error in each case. Hence, a selection procedure was designed to substitute the population distribution produced by number 1, with either 2, 3, 4, or 5 in only those cases where the population distribution produced by number 1 was considered unrepresentative of the estimating year population distribution.

Empirical testing based on the 1981 Census resulted in the following selection procedure to be adopted.

First, all migration distributions possible were calculated and added to the natural base population, resulting in up to five possibilities for the small area estimated population by sex and single year of age in the next time period. These age/sex specific population estimates were then examined to determine which one produced the least change in the small area age structure from the previous year. This was done by first calculating the unweighted average percent difference between the age structures for each of the five possible populations in time $t+1$ to the population in time t . Next, the standard deviations about these averages were calculated, and the distribution with the lowest standard deviation is flagged. If the standard deviation produced by using the Historical Small Area Distribution was significantly greater than the smallest standard deviation (i.e. of the flagged distribution), then the Historical Small Area Distribution was rejected. This procedure was repeated with the Historical Grouped Distribution, and so on until one of the five possible populations was selected.

Once the "best" population in time $t+1$ was calculated for all small areas, two final adjustments were made. First, family allowance data was substituted for the age groups 0-14, and the populations for the rest of the age groups were pro-rated to keep the total population of each small area constant. The second adjustment was to pro-rate the population to ensure the age distribution of the sum of the small area population estimates was consistent with the British Columbia age distribution estimated by Statistics Canada.

5. EVALUATION OF THE CURRENT METHODOLOGY

The following tables summarize the error associated with the June 1, 1981 population estimates by five year age group to 70+, for 74 British Columbia school districts and 29 census divisions. The census division age/sex specific population estimates were derived by aggregating school district population estimates.

The accuracy of the small area age/sex specific population estimates derived from the previously described methodology was evaluated by producing 1981 population estimates by sex and 5 year age groups to 70+ for 74 school districts, then comparing these results to the 1981 Census. Two summary measures were used to evaluate the effectiveness of the age/sex specific population estimates. These were Average Absolute Percent Error (AAPE) and Index of Misallocation (IM). The AAPE is defined as:

$$AAPE = 100 \times \left[\sum_{i=1}^N \left| (P_{Ei} - P_{Ai}) / P_{Ai} \right| \right] / N$$

where P_{Ei} is the estimated cell population for age group i , P_{Ai} is Census cell population for age group i , and N the number of cells. The IM is defined as:

$$IM = 100 \times \frac{1}{2} \left[\sum_{i=1}^N (|P_{Ai} - P_{Ei}|) \right] / \sum_{i=1}^N P_{Ai}$$

where P_{Ai} is the actual cell population for age group i , and P_{Ei} is the estimated cell population for age group i .

As seen in Table 1, relative to the 1981 Census the average absolute percent error over all age groups and regions is 6.20%, and the IM is 1.95%. The average percent errors for male and female are quite similar (AAPE's of 7.00% for both, and IM 's of 2.15% for males and 2.08% for females).

By age, the highest errors occur in the 20-29 and 60-69 age groups. It should also be noted that there is some difference in the age distribution of errors between males and females. Males appear to have higher error in the upper age groups, while females have higher error in the very mobile 20-29 age groups.

Table 1
Error by Age Group Across School District
1981 Estimated Versus Census
Absolute Average Percent Error (AAPE) and Index of Misallocation (IM)

Age	Total		Male		Female	
	AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)
0-4	3.33	0.96	3.94	1.21	3.62	1.04
5-9	2.80	0.76	3.28	0.88	3.62	1.02
10-14	2.33	0.64	3.54	0.84	2.88	0.87
15-19	5.20	2.01	5.68	2.01	6.18	2.24
20-24	13.32	4.77	13.50	4.62	14.54	5.12
25-29	8.31	4.07	8.42	3.70	9.41	4.65
30-34	5.02	2.12	5.42	2.45	5.72	2.06
35-39	4.88	1.33	5.73	1.62	5.38	1.34
40-44	4.52	1.33	5.84	1.51	4.67	1.52
45-49	3.60	1.22	4.47	1.37	4.78	1.49
50-54	5.66	1.33	5.86	1.48	6.68	1.54
55-59	6.11	1.72	6.19	1.78	7.82	1.97
60-64	8.86	2.44	10.35	2.95	8.91	2.17
65-69	10.60	2.66	12.53	3.52	11.44	2.30
70+	8.49	1.95	10.19	2.35	9.33	1.94
Average	6.20	1.95	7.00	2.15	7.00	2.08

As seen in Table 2, on average higher percent errors are associated with areas of small population size. The higher percent errors in smaller areas may be associated with the instability of the smaller (resource based) economies, and associated instabilities in net migration distributions.

By census division, similar error patterns are observed. As seen in Table 3, the average absolute percent error across all regions and age groups is 4.83%, 5.19% for males and 5.60% for females. The IM is 1.27% for the total, 1.41% for males and 1.35% for females. Again, the error is bimodal, with peaks at 20-29 and 60-69. In addition, the females have higher errors than males in the 20-29 age groups, while the reverse is true in the 60-69 age groups.

Table 2
School District Error by Population Size

Population Grouping	Total		Male		Female	
	AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)
0-9,999	8.87	3.16	10.14	3.89	10.27	3.65
10,000-24,999	6.07	2.47	6.92	2.96	6.62	2.58
25,000+	3.66	1.67	3.92	1.78	4.09	1.78
School District Average	6.20	1.95	7.00	2.15	7.00	2.08

Table 3
Error by Age Group Across Census Division
1981 Estimated Versus Census

Age Group	Total		Male		Female	
	AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)
0-4	2.37	0.54	3.20	0.76	2.28	0.58
5-9	1.52	0.50	1.71	0.55	2.13	0.68
10-14	1.69	0.39	2.75	0.57	2.50	0.60
15-19	3.81	1.39	3.79	1.30	4.68	1.63
20-24	9.83	3.07	9.30	2.91	10.90	3.41
25-29	7.02	3.04	7.30	2.87	8.09	3.37
30-34	3.28	1.29	3.31	1.43	3.85	1.25
35-39	3.34	0.66	3.06	0.57	4.21	0.88
40-44	3.86	0.88	4.29	1.01	4.16	0.90
45-49	2.91	0.70	3.20	0.75	3.75	0.83
50-54	4.82	0.64	4.41	0.75	6.10	0.86
55-59	5.49	1.34	5.36	1.55	6.94	1.30
60-64	7.88	1.95	8.37	2.29	7.94	1.74
65-69	8.48	1.89	10.30	2.67	9.79	1.43
70+	6.16	0.81	7.46	1.20	6.73	0.71
Avg	4.83	1.27	5.19	1.41	5.60	1.35

Table 4 (Census Division Error By Population Size) shows the improvement in error levels resulting from aggregating to larger sub-provincial areas. Table 7 illustrates the negative relationship between error levels and population size on a Census Division level.

A comparison of Tables 5 and 6 again demonstrates the improvement in error levels when aggregating to larger age/sex cell sizes. Although this does indicate that some precautions should be observed when utilizing age/sex estimates for some small areas, we do not believe it should preclude use of the estimates for these areas.

Table 4
Census Division Error by Population Size

Population Grouping	Total		Male		Female		N
	AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)	
0-39,000	7.22	1.94	7.55	2.13	8.79	2.29	10
40,000-59,999	4.32	1.82	5.03	2.14	4.91	1.83	10
60,000+	2.51	0.87	2.73	.98	2.84	0.90	9
Census Division Average	4.83	1.27	5.19	1.41	5.60	1.35	29

Table 5
- School District -
Number of Estimates by Error Range

	Average Absolute Percent Error Range				Total
	< 5	5 to 10	10 to 15	15 +	
No. of Cells	674	239	101	96	1110
Percent	61%	22%	9%	9%	100%

Table 6
- Census Division -
Number of Estimates by Error Range

	Average Absolute Percent Error Range				Total
	< 5	5 to 10	10 to 15	15 +	
No. of Cells	306	77	25	27	435
Percent	70%	18%	6%	6%	100%

6. FINAL REMARKS

The procedure outlined above has particular advantages for use in a region with well developed sources of historical small area population and vital statistics data. It is felt that a procedure utilizing net-migration estimates is relatively straightforward, produces acceptable error levels, and can produce age/sex estimates soon after the reference date. Although the optimal situation would be to have in- and out-migration estimates, currently little information is available on small area migration flows within British Columbia. One further improvement to the system being considered is the incorporation of Old Age Security counts to increase the stability and accuracy of estimates in the older age groups.

ACKNOWLEDGEMENTS

The authors would like to thank Don McRae, Steve Miller, Ravi Verma, Garnett Picot, and Paul Knapp whose input to and support for the development of the estimation breakdown system should not go unrecognized.

Table 7
Error by Census Division Across Age Groups
1981 Estimated Versus Census

Census Division	Total Population	Total		Male		Female	
		AAPE (%)	IM (%)	AAPE (%)	IM (%)	AAPE (%)	IM (%)
1000 East Kootenay	53,725	4.24	2.04	5.24	2.29	3.88	2.15
3000 Central Kootenay	52,045	4.00	2.18	4.03	2.13	5.06	1.69
5000 Kootenay-Boundary	33,235	2.32	1.23	2.34	1.18	3.21	1.68
7000 Okanagan-Similkameen	57,185	5.04	2.64	6.02	3.08	4.72	2.49
9000 Fraser-Cheem	56,930	3.12	1.60	3.33	1.78	4.15	2.08
11000 Central Fraser Valley	115,015	3.14	1.43	3.46	1.52	3.65	1.81
13000 Dowdney-Alouette	62,000	2.10	1.15	2.56	1.23	2.26	1.32
15000 Greater Vancouver	1,168,700	1.63	0.94	1.68	0.93	1.67	0.98
17000 Capital	249,475	1.64	0.87	2.31	1.21	1.18	0.61
19000 Cowichan Valley	45,315	3.09	1.66	3.36	1.69	3.85	2.08
21000 Nanaimo	84,815	3.07	1.58	3.40	1.74	3.22	1.66
23000 Alberni-Clayoquot	32,560	2.75	1.36	2.88	1.27	3.27	1.68
25000 Comox-Strathcona	68,620	1.44	0.80	1.85	0.87	2.85	1.50
27000 Powell River	19,050	5.36	2.58	5.06	2.44	6.18	3.03
29000 Sunshine Coast	16,625	4.84	2.57	6.79	3.58	5.65	2.81
31000 Squamish-Lillooet	18,925	1.82	0.99	2.56	1.37	3.10	1.58
33000 Thompson-Nicola	102,430	2.13	1.10	2.07	0.10	2.65	1.37
35000 Central Okanagan	85,235	3.96	1.93	3.91	1.88	4.32	2.14
37000 North Okanagan	69,033	5.26	2.52	6.44	3.06	5.05	2.50
39000 Columbia-Shuswap	45,425	3.04	1.63	3.56	1.84	2.99	1.66
41000 Cariboo	58,810	3.18	1.93	3.90	2.18	3.42	2.06
43000 Mount Waddington	14,675	8.96	3.04	5.13	1.59	17.77	5.49
45000 Central Coast	3,050	17.99	7.62	21.62	8.86	14.92	7.34
47000 Skeena-Queen Charlotte	24,030	4.82	2.09	5.70	2.58	4.61	1.84
49000 Kitimat-Stikina	41,790	6.26	1.99	4.99	1.66	8.59	2.78
51000 Bulkley-Nechako	38,310	6.23	2.31	5.76	2.10	6.83	2.57
53000 Fraser-Fort George	89,430	3.50	1.41	3.39	1.25	3.72	1.68
55000 Peace River-Liard	55,340	8.00	2.95	9.43	3.65	7.34	2.83
57000 Stikine	2,685	17.15	6.89	17.89	6.88	22.39	8.35
Average Error		4.83	2.17	5.19	2.31	5.60	2.51

REFERENCES

- CENTRAL STATISTICS BUREAU (1982). British Columbia municipal population estimation methodology. Unpublished report. Victoria: British Columbia Ministry of Industry and Small Business Development.
- SPSS INC. (1984). *Statistical Package for the Social Sciences/PC*. Chicago, B265-B280.
- SHRYOCK, H.S., and SIEGAL, J.S. (1980). The methods and materials of demography, 2. U.S. Bureau of the Census, 628-630.

Estimating Population by Age and Sex for Census Divisions and Census Metropolitan Areas¹

RAVI B.P. VERMA, K.G. BASAVARAJAPPA and
ROSEMARY K. BENDER²

ABSTRACT

A methodology has been developed for producing population estimates by single years of age and sex for small areas (census divisions and census metropolitan areas). To assure reliability, the estimates by single years of age are grouped into five years and only these grouped data are recommended for dissemination. They are based on the age-sex composition of population from the last census, births by sex, deaths by single years of age and sex, estimates of migration by age and sex, and counts of family allowance recipients in the age group 1-14 years.

KEY WORDS: Cohort-component method; Mean absolute error; Index of dissimilarity; Separation factor.

1. INTRODUCTION

The objective of this paper is to describe the methodology for estimating population by age and sex for small areas (census divisions and census metropolitan areas), present findings of the evaluation of estimation methods, and finally to discuss the factors affecting the quality of estimates. According to the 1981 Census, the 266 Census divisions ranged in population from 2,000 to 2,000,000, and the 24 census metropolitan areas, from 100,000 to 3,000,000. The description of the estimation methods and principal data sources are presented in section 2. The results of the evaluation of migration and population estimates are given in section 3.

2. METHODOLOGY

The descriptions of the estimation methods, as well as the preparation of the basic input data are presented below.

2.1 Cohort-Component Method

For each census division (CD) and census metropolitan area (CMA), the cohort-component method is used to produce population estimates by age. The equations are as follows:

$$\text{For the age 0, } P_0^{t+1} = B - f_0 D_0 + \frac{1}{2} M_0 \quad (1)$$

$$\text{For the age 1, } P_1^{t+1} = P_0^t - [(1-f_0)D_0 + \frac{1}{2} D_1] + \frac{1}{2} (M_0 + M_1) \quad (2)$$

$$\text{For ages 2 to 84, } P_{a+1}^{t+1} = P_a^t - \frac{1}{2} (D_a + D_{a+1}) + \frac{1}{2} (M_a + M_{a+1}) \quad (3)$$

$$\text{For ages 85+, } P_{85+}^{t+1} = P_{84+}^t - \frac{1}{2} D_{84} - D_{85+} + \frac{1}{2} M_{84} + M_{85+} \quad (4)$$

¹ Revised version of the paper presented at the Federal-Provincial Committee on Demography meetings held on November 28-29, 1985 at Statistics Canada, Ottawa, Canada. This research was undertaken with support from the Small Area Data Program of Statistics Canada.

² Ravi B.P. Verma, K.G. Basavarajappa and Rosemary K. Bender, Demography Division, Statistics Canada,, 4th floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

where f_0 = Separation factor of deaths at age 0

M_a = Net migrants aged a between time t and $t+1$

B = Births between time t and $t+1$

D_a = Deaths at age a between time t and $t+1$

P'_a = Population aged a at time t .

The cohort-component method is also used at the provincial level by Statistics Canada (Statistics Canada, Catalogue No. 91-210), and by the province of British Columbia for producing population estimates by age at the census division, school and health district levels (Central Statistics Bureau 1980).

2.2 Preparation of Basic Input Data

Since we are proposing to produce preliminary postcensal population estimates within eight months after the reference date, final data on components of population change cannot be used because they do not become available until after 18 to 24 months. Consequently, estimates would have to be used for each component.

Births and Deaths

Preliminary estimates of births by sex for year (t) are obtained by multiplying the proportional distribution by small areas of provincial total births by sex for year ($t-1$) with the provincial preliminary total births for year (t). Similarly, preliminary estimates of deaths by age and sex for year (t) are obtained by multiplying the proportional distribution by small areas of provincial total deaths by age and sex for year ($t-1$) with the provincial preliminary total deaths for year (t). Finally, they are converted into cohort deaths on the assumption that dates of birth of those who die and the number of deaths are uniformly distributed over a 12 month period except for deaths of age 0. The formulae are as follows:

For age 0,

$$\text{Cohort deaths (0)} = \text{deaths (0)} \times 0.89$$

For age 1,

$$\text{Cohort deaths (1)} = [\text{deaths (0)} \times 0.11] + [\text{deaths (1)} \times 0.5]$$

For ages 2 to 84,

$$\text{Cohort deaths (age)} = [\text{deaths (age-1)} \times 0.5] + [\text{deaths (age)} \times 0.5]$$

$$\text{Cohort deaths (85+)} = \text{deaths (84)} \times 0.5 + \text{deaths (85+)}.$$

In the above formulae, the separation factors (f) are 0.89 for age 0, 0.11 for age 1 and 0.5 for all other ages.

Residual Net Migration

First, the estimates of total population for the postcensal years for CDs and CMAs prepared by the regression-nested procedure are split by sex using the sex composition from the latest census. The regression-nested procedure is described elsewhere (Statistics Canada, Catalogue No. 91-211). For males and females, residual total net migration is computed by taking the difference between the population change and the natural increase. For each area, this is distributed by five year age groups using migration data by age from three sources: residual

net migration from the 1976 and 1981 censuses, migration data from income tax files and the 1981 mobility question. The mobility question referred to is "Where were you on June 1, 1976?" in the 1981 Census. From the responses obtained for this question, in-migrants to and out-migrants from each small area can be tabulated. The five year age groups are split into single years of age using SPRAGUE multipliers. Before applying Sprague multipliers, the residual net migration is first split into in and out migration. Using in and out tax migration data as a reference, this calculation is done individually for each five-year age group.

$$\text{Residual In-Migration} = \frac{\text{Tax Data In-Migration}}{\text{Tax Data Net Migration}} \times \text{Residual Net Migration}$$

$$\text{Residual Out-Migration} = \text{Residual In-Migration} - \text{Residual Net Migration}$$

Using the preceding ratios, major problems occur when the split net migration is not of the same sign as the reference tax data on net migration. In this case, the sign of the split net migration is kept, but the resulting in and out migration are exchanged to yield the appropriate sign. This is based on the assumption of equal magnitude of a reversal of the migration flow.

2.3 Counts From The Family Allowances File, Ages 1-14 years

Estimates of population produced by the cohort-component method for the age groups 1-4, 5-9, 10-14 are replaced by counts of family allowance recipients at these ages which are readily available for CDs and CMAs, within 3 to 4 months after the reference date. Family allowances are paid universally in Canada and hence the counts are considered to be complete for all practical purposes. The data on the family allowance recipients are not provided by sex. Hence they are split into males and females using the sex composition from the latest census.

2.4 Adjustments for Consistency with Provincial and Census Division Estimates

Postcensal regression-nested estimate of total population of each CD and CMA become available within six months after the reference date. In addition, provincial estimates of population also become available by age and sex about the same time. Estimates of population by age and sex prepared as described above for the CDs within each province are controlled with respect to the census division total population estimates, and to the provincial population estimates by age and sex on a pro rata basis. For the census metropolitan areas, the age and sex totals are adjusted only to the CMA total population estimate.

3. EVALUATION

The evaluation is done with respect to three criteria: (i) accuracy; (ii) timeliness and (iii) consistency. Each of these is discussed below.

3.1 Accuracy

The accuracy of population estimates by age and sex depends to a large extent on the accuracy of estimation of the age-sex distribution of migrants, as the data on deaths by age and sex are considered satisfactory. Thus an evaluation of population estimates by age and sex indirectly throws light on the accuracy of migration estimates by age and sex. The accuracy is examined by comparing the estimates with the corresponding census counts.

Table 1
 Distribution of Census Divisions/CMA's Showing the Accuracy
 of Population Estimates by Age, 1981

Provinces	Methods of Migration Estimation	Levels of Mean Absolute Error (%) by Sex							
		Males				Females			
		Under 3	3-5	5-10	10+	Under 3	3-5	5-10	10+
Newfoundland	R	8	2	0	0	10	0	0	0
	M	2	7	1	0	3	3	3	1
	T	0	0	5	5	1	1	3	5
Prince Edward Island	R	1	1	1	0	2	0	0	1
	M	1	2	0	0	1	2	0	0
	T	0	0	1	2	0	0	2	1
Nova Scotia	R	8	4	3	3	8	5	3	2
	M	3	6	4	5	5	6	2	5
	T	0	2	8	8	1	3	11	3
New Brunswick	R	10	1	2	2	7	4	3	1
	M	4	7	3	1	3	8	3	1
	T	0	2	5	8	0	4	5	6
Quebec	R	13	27	23	13	24	18	19	15
	M	12	26	26	12	17	23	21	15
	T	1	5	37	33	3	13	32	28
Ontario	R	30	8	8	7	37	5	4	7
	M	8	16	18	11	21	10	10	12
	T	0	8	34	11	4	10	31	8
Manitoba	R	4	6	8	5	4	6	8	5
	M	1	5	12	5	0	6	7	10
	T	0	1	8	14	0	1	4	18
Saskatchewan	R	10	5	1	2	9	5	2	2
	M	1	11	5	1	4	5	5	4
	T	1	1	10	6	1	1	12	4
Alberta	R	9	3	1	2	8	3	3	1
	M	5	5	3	2	5	4	5	1
	T	0	2	5	8	0	3	5	7
British Columbia	R	19	3	3	4	23	1	1	4
	M	9	13	2	5	14	8	3	4
	T	0	0	13	16	0	3	16	10
CMA	R	14	8	2	0	19	3	2	0
	M	2	17	5	0	10	9	4	1
	T	1	7	13	3	1	10	12	1

Note: R: Residual based age distribution of migrants, 1976-81.

M: Mobility based age distribution of migrants, 1981.

T: Annual tax migration data.

Source: Demography Division, Statistics Canada, 1985.

For each CD and CMA, three sets of population estimates by age and sex as of June 1, 1981 produced by using the age distribution of migrants from the three sources (residual (1976-81), mobility (1976-1981) and annual tax files) and counts from family allowance files as described in sections 2.1 to 2.4 were compared with the 1981 census counts. The differences were termed errors and for each small area, a summary index known as the "mean absolute error" (MAE) was computed by taking the arithmetic mean of percentage errors disregarding

the sign for 16 five year age groups. The smaller the value of this index, the more accurate are the estimates. In Table 1, a classification of CDs by provinces and of CMAs is presented for four levels of mean absolute error: under 3%, 3-5%, 5-10% and over 10%. Overall, it appears that the residual based age distribution of migrants gives better estimates. For males, about 66% of the total number of census divisions had an MAE under 5%. For females this percentage was slightly higher, at 69%. In contrast, lower percentages were observed for the mobility (55% and 57%) and tax migration data (9% and 19%), for males and females, respectively.

For CMAs too, the residual age distribution of migrants seems to give better estimates. The proportions of cases with MAE under 3% were 58% and 79% for males and females respectively. Mobility and tax based age distributions of migrants ranked second and third respectively, for both males and females.

With the exception of Prince Edward Island, the relative accuracy of the three sets of age distribution observed for Canada largely holds good for each province. This is true for both males and females. However, in some cases the residual based age distributions seem to give results similar to the mobility based distributions. Such similarity was observed for males in three provinces (Newfoundland, New Brunswick and British Columbia), whereas for females it was found only in New Brunswick.

It should be noted that the age distribution of migrants derived by the residual method uses the census age distributions of 1976 and 1981. Consequently, the population estimates as of June 3, 1981 prepared by using the migrant age distribution based on the residual method can be expected to be similar to the 1981 census age distribution. Hence, on the basis of this comparison we cannot conclude that the migrant age distribution derived by the residual method is better than the distribution derived from mobility question or from tax files.

Table 2 presents the percentage distribution of CD and CMA outliers. The outliers are those CDs with an MAE of over 10% and those CMAs over 5%. They are presented by sex and the three sources of migrant age distributions. As expected, both for males and females, the proportion of outliers is generally low for estimates using residual based age distribution. On the other hand, the percentage of outliers tends to be high for estimates using tax based migration distribution.

Temporal Stability of the Three Sets of Estimates During Postcensal Years, 1982-1984

For postcensal years, as there are no standard age distributions with which the estimates can be compared, the three population estimates by age and sex are compared with each other to learn of the temporal stability among them. A summary index known as the "index of dissimilarity" calculated as half of the sum of absolute differences in two percentage age distributions is used for this purpose. The range of the index is from 0 to 100. The smaller the value, the greater is the similarity between the two distributions compared. The small areas are classified into three levels of dissimilarity: (i) the smallest level of difference with indices between 0% and 5%; (ii) the medium level of difference with indices between 5 to 10% and (iii) the outliers showing the index value of 10% and over. The classification of CDs is presented in Table 3 and that of CMAs in Table 4.

From Table 3, it appears that all the three population distributions tend to be similar and on average, a high percentage of cases, about 90%, are in the smallest category of differences (0%-5%) with only about 7% falling in the 5% to 10% category.

The percentage of cases with the extreme level of differences (index of dissimilarity exceeding 10%) were also examined for the ten provinces and their total. For males, the percentages of extreme cases were small, 3 to 5% between the residual and mobility based age distributions. For females, a relatively higher proportions of outliers were noticed. For other comparisons, residual vs tax based, and mobility vs tax based, slightly higher proportions of outliers were found. The results were similar for census metropolitan areas (see Table 4).

Table 2Percentage of Outliers^a Among Census Divisions by Province, and of CMA's 1981

Provinces	Males			Females		
	R	M	T	R	M	T
Newfoundland	0	0	50	0	10	50
Prince Edward Island	0	0	67	33	0	33
Nova Scotia	17	28	44	11	28	17
New Brunswick	13	7	53	7	7	40
Quebec	17	16	43	20	20	37
Ontario	13	21	21	13	23	15
Manitoba	22	22	61	22	43	78
Saskatchewan	11	6	33	11	22	22
Alberta	13	13	53	7	7	47
British Columbia	14	17	55	14	14	34
Total	15	16	43	15	20	35
CMA	8	21	67	8	21	54

Note: R: Residual based age distribution of migrants.

M: Mobility based age distribution of migrants.

T: Tax based age distribution of migrants.

^a The outliers are those CDs with MAE of over 10% and those CMAs with MAE of over 5%.

Source: Table 1.

Table 3Distribution of Census Divisions by Level of Index of Dissimilarity
Obtained by Comparing the Age Distributions of Population Based on Residual,
Mobility and Tax Migration Sources, 1982 to 1984

Year/ Index of Dissimilarity	Males			Females		
	Residual vs Mobility	Residual vs Tax	Mobility vs Tax	Residual vs Mobility	Residual vs Tax	Mobility vs Tax
YEAR 1982						
0-5	245	237	242	240	241	234
5-10	7	13	10	8	5	11
10+	8	10	8	12	14	15
Total	260	260	260	260	260	260
YEAR 1983						
0-5	235	221	223	230	229	223
5-10	11	18	21	10	13	13
10+	14	21	16	20	18	24
Total	260	260	260	260	260	260
YEAR 1984						
0-5	240	226	229	235	233	231
5-10	11	16	14	15	13	12
10+	9	18	17	10	14	17
Total	260	260	260	260	260	260

Source: Demography Division, Statistics Canada, October 1985.

Table 4

Distribution of Census Metropolitan Areas by Level of Index of Dissimilarity
Obtained by Comparing the Age Distributions of Population Based on Residual,
Mobility and Tax Migration Sources, 1982 to 1984

Year/ Index of Dissimilarity	Males			Females		
	Residual vs Mobility	Residual vs Tax	Mobility vs Tax	Residual vs Mobility	Residual vs Tax	Mobility vs Tax
YEAR 1982						
0-3	24	24	24	23	22	23
3-5	0	0	0	0	1	0
5+	0	0	0	1	1	1
Total	24	24	24	24	24	24
YEAR 1983						
0-3	22	23	22	21	20	20
3-5	2	0	0	1	2	3
5+	0	1	2	2	2	1
Total	24	24	24	24	24	24
YEAR 1984						
0-3	21	21	21	21	20	20
3-5	2	2	2	0	1	0
5+	1	1	1	3	3	4
Total	24	24	24	24	24	24

Source: Demography Division, Statistics Canada, October 1985.

In conclusion, it may be said that although the three age distributions of migrants (residual, mobility and tax based) differed from each other, age distributions of population resulting from these were largely similar.

3.2 Timeliness

Timeliness refers to the availability of estimates within as short a time as possible after the reference date. Using the preliminary population totals (regression-nested estimates) which become available within six months from the reference date, the estimated numbers of births, deaths by age and net migrants by age as described in Sections 2.1 to 2.4, the population estimates by age and sex for CDs and CMAs could be prepared within eight months of the reference date.

3.3 Consistency

Consistency refers to the consistency in the sources of data sets used for estimation at various levels of administrative or other disaggregated areas and to the uniformity in the methods of estimation. While in certain cases, a different method may have to be used, it is highly desirable to use the same method throughout in order to ensure the methodological consistency of various levels of geographic disaggregation.

For provinces, CDs and CMAs, the sources of data are the same for births and deaths: the vital registration records. For migration data too, the sources are the same namely, tax files and mobility data from the census for all levels of geographic disaggregation. However, an additional data set, the residual age data derived from the two consecutive censuses is also used.

There is full methodological consistency between provinces and other levels as the cohort-component method is used in all cases.

4. CONCLUSION

By using the cohort-component method, three sets of estimates by age and sex have been prepared for CDs and CMAs. Each set uses a different migration component by age and sex: (i) tax file based; (ii) mobility data from the latest census and (iii) the residual derived from the two consecutive censuses.

Although the three age distributions of migrants differ from each other, the resulting estimates of population by age and sex were largely similar. Each set involves its own assumptions. Using a residual age distribution of migrants for postcensal estimation assumes that the age distribution remains constant for the period of estimation. A similar assumption is involved in using mobility data by age and sex for postcensal years. The data from tax files assume that the age-sex distribution remains the same for any two consecutive years. However, the type of movement measured by each of these sets is not the same. The residual measures only the net movement between the two consecutive censuses (e.g. 1976-81). The mobility question also measures five-year movements ranging from 0-4 years. The tax files, on the other hand show the movement during roughly a 12 month period. On the basis of the comparisons made in the paper, it cannot be concluded that one migrant data set giving rise to population estimates is better than the other. A more satisfactory evaluation of the three sets of estimates can be made only when the next census results become available.

REFERENCES

- CENTRAL STATISTICS BUREAU (1980). Population estimates by age groups for school districts, 1977-80. Unpublished Technical Report, Government of B.C.
- NORRIS, D., and STANDISH, L. (1983). A Technical report on the development of migration data from taxation records. Technical Report, Administrative Data Development Division, Statistics Canada.
- STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population by Marital Status, Age, Sex and Components of Growth for Canada and the Provinces*, Vol. 2, 2nd issue. Catalogue 91-210, Ottawa: Ministry of Supply and Services.
- STATISTICS CANADA (Annual). *Postcensal Annual Estimates of Population for Census Divisions and Census Metropolitan Areas (Regression-Method)*. Catalogue 91-211, Ottawa: Ministry of Supply and Services Canada.
- STONE, Leroy O. (1980). Evaluating the relative accuracy and significance of net migration estimates. *Demography*, 4, 310-330.

Experience with Small Area Population Estimates¹

ROSEMARY K. BENDER²

ABSTRACT

Statistics Canada's current methodologies forestimating the population of census divisions and census metropolitan areas are the regression-nested and component methods. This paper presents the experience with these estimates for the period 1981 to 1985, focusing on problems encountered with the input data on family allowance recipients.

KEY WORDS: Regression-nested estimates; Component estimates; Family allowance recipients; Postal code files.

1. INTRODUCTION

Statistics Canada's current methodologies for estimating the population of census divisions (CDs) and census metropolitan areas (CMAs) are the regression-nested and component methods. The regression estimates for 1982, 1983 and 1985 were published in Catalogue No. 91-211 on schedule. Those for 1984 were only made available in March of 1985. There was a delay in obtaining the input data on family allowance. Furthermore, as explained below, we encountered problems with the quality of these data. In particular, the resulting population estimates for CMAs were not acceptable and an alternate methodology had to be used.

Component estimates of the population for CDs and CMAs have been published in Catalogue No. 91-212 on schedule for 1982 and 1983. We should release the 1984 estimates by April 1986. An evaluation of the component estimates produced thus far has shown the data to be of good quality.

2. ADJUSTMENTS

Since introducing the regression estimates for CDs and CMAs in 1982, some adjustments to the data and the methodology have been necessary. They are summarized below:

- For the 1983 estimates for the CD Chicoutimi and the CMA Chicoutimi- Jonquière in the province of Quebec, the family allowance data was adjusted based on the growth pattern of the previous year. The problem was traced to postal codes used to obtain the family allowance data.
- In 1984, 17 census divisions estimates were imputed with preliminary component estimates.
- In 1984, we decided to publish for the CMA of Calgary, estimates based on the annual census conducted by the city. This will be done for the entire 1981-1986 period.
- In 1984, we developed a new methodology for all CMAs other than Calgary, which aggregates census division regression estimates. This will be used for the entire 1981-1986 period.

The following sections explain the problems encountered in more detail.

¹ Abridged version of the paper presented at the meetings of the Federal-Provincial Committee on Demography held on November 28-29, 1985, Ottawa, Canada.

² Rosemary K. Bender, Demography Division, Census and Demographic Statistics Branch, Statistics Canada, 4th floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

3. PROBLEMS WITH INPUT DATA FOR REGRESSION ESTIMATES

There was a delay in producing 1984 estimates due to problems encountered in obtaining data on family allowance recipients from Health and Welfare Canada, and the appropriate postal code translation files necessary to process these data.

i) *Family Allowance Data*

The numbers of Family Allowance recipients as of June 1, is generally available by mid September of each year. The 1984 data from Health and Welfare Canada however, were delayed as a result of decentralization of the regional operations of the program in Ontario. Problems were also encountered in the files of all provinces with respect to information on effective dates of transfer and reason codes for inter-area transfers. The 1984 data were released to Statistics Canada in an unedited form in November. Corrective actions were taken by Health and Welfare Canada, and Family Allowance data as of June 1, 1985 was on schedule.

ii) *Postal Code Files*

The data on family allowance recipients from Health and Welfare Canada is coded by postal code. Therefore, to identify the children receiving family allowance in each CD and CMA, a file must be created that groups the postal codes by CD and CMA. This is done using a master file that contains all the postal codes in Canada, with detailed geographic codes that are used to assign the postal codes to any level of geographic disaggregation.

Problems have arisen that were unexpected and in some cases had serious consequences. For our estimates, it is important that the postal code files used each year by Health and Welfare Canada be consistent with the one that was used to develop the regression model. The only change in the file should be the addition of new postal codes. Any shifting of postal codes from one region to another can result in changes to the population that do not actually occur.

The problems we encountered stem from the fact that since we developed our regression model, different divisions and departments have produced the postal code files. In 1982 and 1983, it was done by the Administrative Data Development Division of Statistics Canada. In 1984, the Standards Division of Statistics Canada took over the responsibility and in 1985 it was done by Health and Welfare Canada. Each had its own approach resulting in family allowance data that was not consistent from year to year. Two different types of problems arose. We have resolved the first. However, the second will persist throughout the 1981-1986 postcensal period.

The first source of difficulty was the shifting of postal codes from one area to another. The master file is created by the Standards Division of Statistics Canada. However, in some cases, the CD or CMA geographic code is blank or wrong. For CDs this occurs mostly with rural codes, where postal codes often refer to post offices covering large territories across CD boundaries. The inclusion of the CMA geographic codes is fairly recent, and the quality improves each year. Thus, our initial assumption that the postal code file would be consistent from year to year was not quite true. There are changes made each year.

Our files were initially created by the Administrative Data Development Division (ADDD) of Statistics Canada. They made changes in their copy of the master file before proceeding to group the data. In 1984 the Standards Division took over producing our file. When we became aware of the consequences this would have, we developed with ADDD a way to match the original master file with the latest master file from Standards Division, adding only the new postal codes. Any changes to the CD or CMA codes were ignored. We realise that by doing this we do not have the most accurate postal code file available. However, for our

purposes, we are interested in the changes to the proportions of children receiving family allowance. The effect of using some erroneous, but consistent postal codes is that we include or exclude some children from another area in the calculation of proportions. The proportions would not be significantly different from those using correct postal codes, but would change if these children were suddenly excluded or included.

This process of adding only new codes to our postal code file improved significantly the quality of the 1984 family allowance data for census divisions. Only 17 of the 231 regression estimates of CDs (excluding those of British Columbia, as they produce their own regression estimates) needed to be imputed. Because of the delay in obtaining the data, we were able to use preliminary estimates from the component method. For census metropolitan areas, there were still inconsistencies, which we believe are due to a different type of problem.

When the postal codes are grouped by CDs and CMAs, they are also converted into ranges of postal codes. For example, if the postal codes A1A1A1, A1A1A2, A1A1A3 and A1A1A4 all have the same CMA code, then they will be combined into the range A1A1A1-A1A1A4. However, in processing the over 600,000 postal codes, certain assumptions are made, depending on the software. If, in the above example A1A1A2 was not there, the program may still create the same range, assuming that if A1A1A2 did exist, it would have the same CMA code as the others in the range. This type of assumption could alter the family allowance data processed for each region. Furthermore, if different softwares are used each year, serious inconsistencies can arise.

We believe this is the major cause for the poor quality in the family allowance data for CMAs. The softwares used by the ADDD and Standards Divisions were different. What complicated matters even more was that as of 1985, the entire operation is now done by Health and Welfare Canada, again using a different software. We therefore had to disregard the data and develop an alternative methodology for CMAs.

4. METHODOLOGICAL CHANGE FOR CMAs

The CMA estimates previously released for 1982 and 1983 were based on the same regression-nested procedures as for census divisions. In the evaluation of the 1984 estimates, however, estimates for many census metropolitan areas were found to be inconsistent with alternate sources and past growth trends. As described above, the problems seem more related to the quality of the input files rather than to methodology.

Taking into account these inconsistencies as well as comments from the provincial focal points, it was decided to use an alternate methodology. This new methodology was previously developed for estimating various CMA components of population change. It consists of aggregating census divisions regression estimates, using the ratio of the population of the CMA to that of overlapping CDs, as observed the previous year by the component method. In comparing estimates for 1981, obtained through this methodology, with the 1981 Census counts for census metropolitan areas, an average absolute error of 1.3% as observed, as compared to 2.3% for the previous methodology.

To maintain consistency in methodology for the entire 1981-1986 period, the alternate method has been used to derive the CMA estimates for 1982 to 1985, and will be used for 1986. That is, estimates of population for CMA's other than Calgary are obtained by aggregating the census division regression-nested estimates, and those for Calgary as described below, are based on the annual census conducted by the city.

In 1984, it was found that the regression-nested estimates for Calgary CMA for 1982 and 1983 were too high in comparison with the census counts conducted annually by the city of Calgary. The component estimates also supported the idea of adjusting the regression-nested estimates for Calgary. It was decided to publish estimates based on the city of Calgary

census count extrapolating the April data to June 1. This is in line with Statistics Canada policy where, when there is a complete enumeration, this should be considered over an estimate prepared by an indirect procedure, unless there is evidence that the enumerated count is suspect.

5. COMPARISON WITH OTHER DATA SOURCES

The regression and component estimates are compared with alternative data sources whenever possible. We receive from the Saskatchewan and Alberta governments the number of people registered in their respective health care programs. These data are used in the regression model. However, they are also evaluated for consistency with the family allowance data and past growth trends. In most cases they were consistent, and differences were traced to the problems encountered with family allowance data.

The Quebec Bureau of Statistics produces annual population estimates of their administrative regions which are subdivisions of the Quebec CDs. Their data are comparable to ours except for the CD of Nouveau Québec. This census division, located in northern Quebec, is largely comprised of unorganized territories, and it is difficult to estimate the population. The BSQ generally adopts our estimates, though for 1984 it imputed its own estimate for Nouveau Québec.

We also appreciate feedback from users who may have access to specific local area data.

6. CONCLUSION

The methods used to produce population estimates for census divisions and census metropolitan areas have in general functioned very well. However, in the case of the regression estimates, problems with input data made it necessary to impute estimates for certain CDs with alternate data, and to revise the methodology for CMAs.

The problems encountered were mostly related to the family allowance data and the postal code files that are necessary to process these data. Most of the problems have been resolved. However, as Health and Welfare are now taking over the responsibility of creating the postal code files, the 1986 data may still have problems of consistency and will have to be carefully evaluated.

Despite these problems, the regression methodology with certain adaptations will be used to produce estimates for 1986. If, however, we decide to continue with the methodology for the 1986-1991 period, we must first ensure that consistent postal code files be processed by the same department throughout the period.

EDITORIAL COLLABORATORS

Following the Journal's policy, all papers published in Volume 11 (issues 1 and 2) were refereed except the paper by M. Wilk and the papers selected from those presented at the meetings of the Federal-Provincial Committee on Demography held on November 28-29, 1985, Ottawa. (Condensed version of these selected papers are included in this issue to provide the readers with additional information on recent methodological developments in an important area of applications.)

The Survey Methodology Journal wishes to thank the following persons who have served as referees during the past year.

D.A. Binder

Y.P. Chaubey

G.H. Choudhry

E.B. Dagum

J. Gambino

G.B. Gray

M.A. Hidioglou

S.K. McKenzie

M. Morry

D.A. Pierce

B. Quenneville

C.E. Särndal

A. Satin

A. Singh

J. Tourigny

A. Van Barren



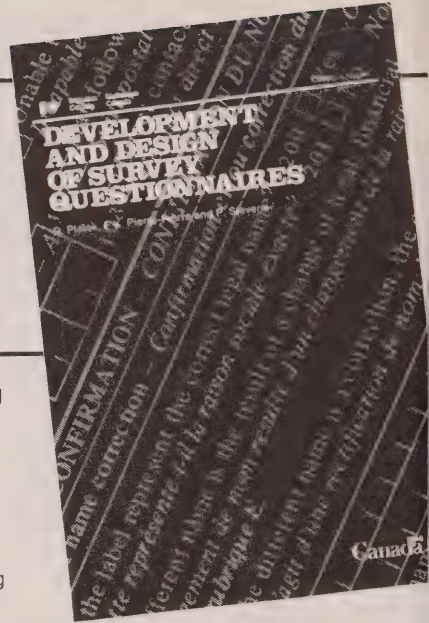
DEVELOPMENT AND DESIGN OF SURVEY QUESTIONNAIRES

Successful implementation of surveys depends to a large extent on good questionnaire design. But the failure to devote sufficient attention, care and resources to questionnaire development is surprisingly common in current survey practice. As a consequence, many surveys fail to achieve their full potential.

From its Introductory comment through its four chapters dealing with organizing to design, development, production and evaluation of questionnaires, *Development and Design of Survey Questionnaires* intends to promote good questionnaire design and serve as a reference and training tool. It provides a discussion of issues related to wording, format, layout, etc., illustrating these with examples from recent federal surveys. A Checklist and Bibliography complete the 119 page text. (15 cm. x 23 cm.)

Contents

- Preface & Introduction
- The Process of Questionnaire Development
- Questionnaire Design
 - Data Quality
 - Grouping Subjects
 - Making Concepts Operational
 - Questionnaires & Schedules
 - Wording of Questions
- Parts of the Questionnaire
 - Open ended & Closed-ended Questions
 - Attitude Scales
- Questionnaire Production
 - Layout
 - Data coding and Capture
 - Administering the Questionnaire
- Testing and Evaluation of the Questionnaire
- Checklist Summary of the Elements of Questionnaire Design
- Bibliography



ORDER FORM

PF 02922

Mail to:
Publications Sales and Services
Statistics Canada
Ottawa, K1A 0T6

(Please print)

Company: _____

Dept.: _____

Attention: _____

Address: _____

City: _____

Tel.: _____

Province: _____

Postal Code: _____

☐ Purchase Order Number (Please enclose) _____

☐ Payment enclosed \$ _____

CHARGE TO MY:

☐ MASTERCARD ☐ VISA ☐ Statistics Canada

Account No.:

Expiry date

☐ Bill me later

My client reference number is: _____

Signature: _____

Catalogue No.	Title	Quantity	Price	Total
12-519E	Development and Design of Survey Questionnaires		\$25 in Canada \$26 other countries	

Cheques or money orders should be made payable to the Receiver General for Canada/Publications, in Canadian funds or equivalent.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, l).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour le première fois.)

5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Exemple: Cochran (1977, p. 164).

LABORATION ET CONCEPTION DES QUESTIONNAIRES ENQUÊTE

accès d'une enquête dépend dans une grande mesure de la conception
questionnaire. Or, il est étonnant de constater qu'on néglige très souvent
l'ordre suffisamment d'attention, de soin et de ressources à l'élaboration
questionnaires d'enquête. En conséquence, bien des enquêtes ne pro-
nt pas tous les résultats qu'on pourrait en attendre.

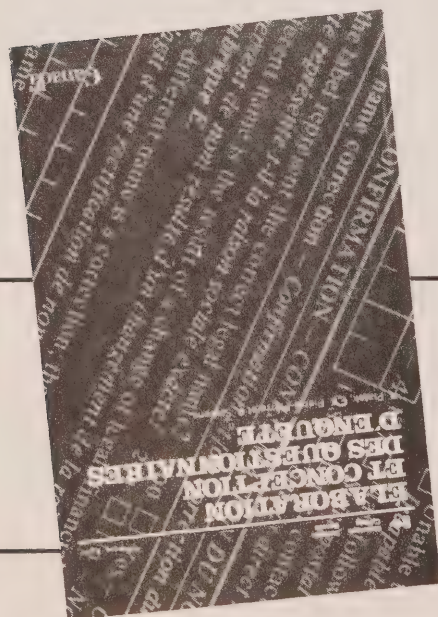
l'introduction et ses quatre chapitres portant sur la conception, l'éla-
et la conception des questionnaires d'enquête, vise à promouvoir une
l'information. On y traite de questions aussi diverses que la formulation
questions, la présentation, la disposition, etc., au moyen d'exemples
l'élites fédérales effectuées récemment. Une liste récapitulative et une
graphie complètent ces 132 pages (15 cm. x 23 cm.).

des matières

roduction
nception du questionnaire
mbinaison de sujets
alité des données
nsposition des concepts
questionnaire et
ormule d'enquête

Formulation des questions
Ordre des parties du questionnaire
Questions ouvertes et questions fermées
● Production du questionnaire
Présentation
Codage et saisie des données
Impression
Administration du questionnaire

- Mise à l'essai et évaluation du questionnaire
- Liste récapitulative des étapes de la conception d'un questionnaire
- Bibliographie



N DE COMMANDE		PF 02923
es moultées s.v.p.)		
at: cations-Vente et service		
tique Canada		
wa, K1A 0T6		
ne: pagne:		
ce: tion:		
se: sse:		
Code postal:		
Titre		
Elaboration et conception des questionnaires d'enquête		
\$25 au Canada		
\$26 autres pays		
Total		
Quantité		
Prix		
Signature:		
N° de compte:		
Date d'expiration		
Facturez-moi plus tard		
Numéro de référence du client:		
N° de la commande (inclure s.v.p.)		
Portez à mon compte:		
Paiement inclus \$		
N° de la commande (inclure s.v.p.)		
Mastercard		
VISA		
Statistique Canada		

COLLABORATION À LA RÉDACTION

En conformité avec la politique rédactionnelle de *Techniques d'enquête*, les articles publiés dans le volume 11 (numéro 1 et 2) ont fait l'objet de critiques, sauf celui de M. Wilk et les exposés présentés à la conférence du comité fédéral-provincial de la démographie, tenue à Ottawa les 28 et 29 novembre 1985. (Nous faisons paraître des versions condensées de ces textes dans un domaine important de leurs applications.)

Les responsables de la revue *Techniques d'enquête* désirent remercier les personnes suivantes qui ont accepté de faire la critique des articles présentés au cours de l'année dernière:

D.A. Binder	M. Morry
Y.P. Chaubey	D.A. Pierce
G.H. Choudhry	B. Quenneville
E.B. Dagum	C.E. Særdal
J. Gambino	A. Satin
G.B. Gray	A. Singh
M.A. Hidiroglou	J. Tourigny
S.K. McKenzie	A. Van Barren

6. CONCLUSION

En règle générale, les méthodes utilisées pour produire les estimations de la population dans le cas des estimations par régression, il a fallu imputer les estimations de certaines DR à l'aide d'autres genres de données et réviser la méthode de calcul des estimations relatives aux RMR, à cause des problèmes d'obtention des données d'entrée.

Les problèmes qui se sont posés avaient trait pour la plupart à l'utilisation des données sur les bénéficiaires d'allocations familiales et des fichiers de codes postaux. Une bonne partie de ces problèmes, toutefois, ont pu être réglés. Par ailleurs, il est possible que, du fait que Santé et Bien-être Canada soit maintenant chargé de créer les fichiers de codes postaux, il y ait encore des cas d'incohérence dans les données de 1986 et qu'il faille procéder à une évaluation rigoureuse.

Malgré ces difficultés, les estimations relatives à 1986 seront produites à l'aide de la méthode de régression qui sera quelque peu remaniée. Si, toutefois, on décidait de continuer d'appliquer les méthodes actuelles pour la période 1986-1991, il faudrait tout d'abord s'assurer que des fichiers de codes postaux compatibles seront traités par le même ministère ou service durant toute la période visée.

d'évaluation des estimations de 1984, les estimations d'un bon nombre de régions métropolitaines de recensement ont été jugées incompatibles avec celles tirées d'autres sources et non conformes aux tendances passées. Comme il a été mentionné plus haut, les difficultés qui ont été soulevées tenaient plus à la qualité des données des fichiers qu'aux méthodes de calcul. Une fois pris en considération le manque d'uniformité des estimations et les observations faites par les points de contact dans les provinces, il a été décidé d'adopter une autre méthode, celle qui avait été conçue pour estimer diverses composantes de l'accroissement démographique de la population. Selon cette méthode, on agrège les estimations par régression relatives aux divisions de recensement, à l'aide du ratio de la population d'une RMR et de celle des DR qui la chevauchent, en fonction des estimations de l'année précédente établies par la méthode des composantes. Par la comparaison des estimations relatives à 1981 calculées selon cette méthode et des chiffres de population des régions métropolitaines de recensement selon le recensement de 1981, on obtient une erreur absolue moyenne de 1.3%, comparativement à 2.3% selon la méthode de calcul des estimations emboîtées.

Pour assurer un degré d'uniformité des données pour la période de 1981-1986, cette nouvelle méthode a été appliquée aux calculs des estimations de la population des RMR pour les années 1982 à 1985, et elle sera utilisée également pour 1986. Ainsi, des estimations de la population des RMR autres que celle de Calgary sont établies par agrégation des estimations emboîtées relatives aux divisions de recensement, et les estimations relatives à Calgary (voir explication plus bas) sont fondées sur les résultats du recensement annuel mené par cette ville. En 1984, les estimations emboîtées relatives à la RMR de Calgary pour 1982 et 1983 ont été trop élevées par rapport aux chiffres du recensement annuel de Calgary. Les estimations calculées par la méthode des composantes ont servi à corriger les estimations emboîtées relatives à cette ville. Il a été décidé de publier des estimations fondées sur les chiffres du recensement de Calgary, en extrapolant les données d'avril au 1 juin. Cela est conforme à la ligne de conduite de Statistique Canada, selon laquelle, lorsqu'il y a un recensement de la population, les chiffres obtenus doivent être utilisés de préférence aux estimations calculées à l'aide d'une méthode indirecte, à moins qu'on n'ait des motifs évidents de douter de la validité des chiffres du recensement.

5. COMPARAISON D'AUTRES SOURCES DE DONNÉES

Chaque fois que cela est possible, les estimations par régression et les estimations calculées par la méthode des composantes sont comparées avec l'information d'autres sources possibles. En effet, la Saskatchewan et l'Alberta nous fournissent des données sur le nombre de bénéficiaires de leur programme de soins de santé; cette information est utilisée dans le modèle de régression. Toutefois, ces estimations font l'objet d'une évaluation qui vise à vérifier leur uniformité par rapport aux données sur les bénéficiaires d'allocations familiales et sur les tendances passées. Dans la plupart des cas, ces estimations étaient compatibles, et les cas de divergences ont pu être attribués aux difficultés liées aux données sur les allocations familiales.

Le Bureau de la statistique du Québec produit des estimations annuelles de la population relatives aux régions administratives, lesquelles sont des subdivisions des DR du Québec. Ces estimations sont comparables à celles de l'Agence fédérale, sauf celles relatives à la DR du Nouveau-Québec. En effet, cette division qui est située dans le nord du Québec comprend essentiellement des territoires non municipalisés, ce qui complique l'estimation de la population. Le B.S.Q. utilise habituellement les estimations de Statistique Canada, quoique, pour 1984, il ait imputé ses estimations pour le Nouveau-Québec.

Statistique Canada est toujours reconnaissant de l'information fournie par les utilisateurs qui disposent de données particulières sur des localités.

Dans certains cas, le code géographique d'une DR ou RMR est inexact ou inexistant. Cette situation se produit la plupart du temps avec les codes ruraux dans les DR, parce que les codes postaux correspondent souvent aux codes d'identification des bureaux de poste qui desservent de grands territoires à travers plus qu'une DR. L'inclusion des codes géographiques des RMR est assez récente et leur fiabilité s'accroît chaque année. Par conséquent, notre première présomption selon laquelle le fichier des codes postaux serait uniforme d'une année à l'autre n'était pas tout à fait juste.

Les premiers fichiers de Statistique Canada ont été créés par la Division de l'exploitation des données administratives (DEDA) qui a modifié sa version du fichier principal avant de procéder au regroupement des données. En 1984, la Division des normes a pris la relève pour la production de ce fichier. Lorsque les conséquences de ce transfert de responsabilité sont devenues évidentes, on a conçu en collaboration avec la DEDA un moyen de faire concorder le fichier principal original avec le dernier fichier principal créé par la Division des normes, en introduisant uniquement les nouveaux codes postaux et en ne tenant pas compte des corrections apportées à des codes DR ou RMR. Il faut admettre qu'en procédant de cette façon on n'obtient pas le fichier de codes postaux le plus exact qui soit. Cependant, pour la production des estimations, il importe de connaître les variations dans les proportions du nombre d'enfants bénéficiaires d'allocations familiales. Le fait d'utiliser certains codes postaux inexacts mais uniformes a pour effet d'inclure des enfants dans une région ou de les exclure d'une autre région de bénéficiaires. Les résultats de ces calculs ne sont pas très différents de ceux qu'on obtiendrait à partir de codes postaux exacts, mais ils seraient différents si on incluait ou excluait subitement ces enfants dans les calculs.

L'introduction uniquement des nouveaux codes postaux dans le fichier de Statistique Canada a amélioré considérablement la qualité des données de 1984 sur les allocations familiales au niveau des divisions de recensement. On a dû imputer les estimations par régions de seulement 17 DR sur 231 (sans les DR de la C.-B., parce que cette province produit ses propres estimations). À cause des retards dans la collecte des données, cela a permis d'utiliser les estimations provisoires calculées à l'aide de la méthode des composantes. Dans le cas des régions métropolitaines de recensement, il y avait encore des incohérences qui sont imputables à un autre genre de problème.

En effet, lorsque les codes postaux sont groupés par DR et RMR, ils sont également répartis par intervalle de codes. Par exemple, si les codes postaux A1A1A1, A1A1A2, A1A1A3 et A1A1A4 correspondent tous au même code de RMR, alors ils sont groupés dans l'intervalle A1A1A1-A1A1A4. Toutefois, lorsqu'on traite la série de plus de 600,000 codes postaux, il faut poser certaines hypothèses selon le logiciel utilisé. Si, dans l'exemple donné plus haut, le code A1A1A2 n'apparaissait pas, le programme créerait quand même le même intervalle, car il supposerait que si le code A1A1A2 existait il aurait le même code de RMR que les autres codes postaux de cet intervalle. Ce genre de supposition pourrait modifier les données sur les allocations familiales traitées relativement à chaque région. En outre, l'application de divers logiciels dans une année pourrait produire des incohérences graves.

À notre avis, il s'agit là de la principale cause de la mauvaise qualité des données sur les bénéficiaires d'allocations familiales au niveau des RMR. Les logiciels que la DEDA et la Division des normes utilisent sont différents et, ce qui n'arrange pas les choses, depuis 1985, le programme relève de Santé et Bien-être Canada qui utilise lui aussi un autre logiciel. Il faut donc laisser de côté ce genre de données et élaborer une autre méthode d'estimation de la population des RMR.

4. NOUVELLE MÉTHODE POUR LES RMR

Les estimations de la population des RMR pour 1982 et 1983 étaient des estimations emboîtées du genre de celles relatives aux divisions de recensement. Dans le programme

- En 1984 également, une nouvelle méthode a été mise au point pour toutes les RMR hors celle de Calgary; cette méthode permet d'agréger les estimations par régression relatives aux divisions de recensement. Ce projet sera appliqué pour toute la période 1981-1986 visée.

Les difficultés que la production de ces estimations a soulevées sont exposées en détail dans les prochaines sections.

3. COLLECTE DES DONNÉES NÉCESSAIRES AU CALCUL DES ESTIMATIONS PAR RÉGRESSION

Il y a eu des retards dans la production des estimations de 1984 à cause de la difficulté d'obtenir les données sur les bénéficiaires d'allocations familiales auprès de Santé et Bien-être Canada, et d'utiliser les bons fichiers de conversion des codes postaux qui sont essentiels au traitement de ces données.

i) *Données sur les allocations familiales*

Le nombre de bénéficiaires d'allocations familiales, en date du 1 juin, est habituellement connu à la mi-septembre de chaque année. Cependant, les données de 1984 n'ont pas été fournies à temps par Santé et Bien-être Canada à cause de la décentralisation des opérations régionales du programme en Ontario. Il y a eu également des problèmes d'extraction, dans les fichiers provinciaux, de données sur la date d'entrée en vigueur des transferts, de même que des codes de raison des transferts interrégionaux. Les données de 1984 fournies à Statistique Canada ont été présentées sous une forme non épurée, en novembre. Des mesures correctives ont été prises par Santé et Bien-être Canada, et les données sur les allocations familiales au 1^{er} juin 1985 ont été diffusées comme prévu.

ii) *Fichiers des codes postaux*

Un code postal est attribué à chaque donnée sur les bénéficiaires d'allocations familiales recueillies auprès de Santé et Bien-être Canada. Aussi, pour dénombrer les enfants bénéficiaires d'allocations familiales dans chaque DR et RMR, il faut créer un fichier qui regroupe les codes postaux par DR et RMR. Pour cela il faut utiliser un fichier principal qui contient tous les codes postaux au Canada, de même que des codes géographiques précis qui permettent d'attribuer les codes postaux selon chaque niveau de répartition géographique.

Des difficultés imprévues se sont produites dans l'utilisation des codes et, dans certains cas, elles ont eu des conséquences graves. Pour la production de ces estimations, il est important que les fichiers de codes postaux que Santé et Bien-être Canada utilise chaque année soient compatibles avec celui que Statistique Canada a construit en fonction du modèle de régression. Le seul changement qu'on devrait apporter à ce fichier est l'ajout de nouveaux codes postaux. Le transfert de codes postaux d'une région à une autre peut entraîner des modifications de la composition de la population qui ne se sont pas vraiment produites.

Aussi, les problèmes qui se sont posés viennent du fait que, depuis la construction du modèle de régression de Statistique Canada, divers ministères et divisions ont créé des fichiers de codes postaux. En 1982 et 1983, c'est la Division de l'exploitation des données administratives (DEDA) de Statistique Canada qui l'a fait; l'année suivante, la Division des normes de Statistique Canada a été chargée de cette tâche et, en 1985, c'est le ministère de la Santé et du Bien-être qui a pris la relève. Chacun a adopté une méthode particulière pour constituer ces fichiers, de sorte que les données sur les allocations familiales n'étaient pas uniformes d'une année à l'autre. Cette situation a créé deux genres de problèmes, dont un a été résolu. Le second, toutefois, persistera pendant toute la période postcensitaire de 1981 à 1986.

La première cause de difficultés a été le transfert de codes postaux d'une région à une autre. Le fichier principal des codes a été créé par la Division des normes de Statistique Canada.

Estimations de la population des petites régions: expérience de Statistique Canada¹

ROSEMARY K. BENDER²

RÉSUMÉ

À Statistique Canada, les estimations de la population des divisions et régions métropolitaines de recensement sont calculées à l'aide de la méthode des composantes et d'un modèle de régression qui produit des estimations emboîtées. Le présent document décrit les travaux effectués dans ce domaine relativement à la période 1981-1985 et expose particulièrement les difficultés qu'a posées l'utilisation des données sur les bénéficiaires d'allocations familiales.

MOTS CLÉS: Estimations emboîtées; estimations de la méthode des composantes; bénéficiaires d'allocations familiales; fichiers de codes postaux.

1. INTRODUCTION

Pour estimer la population des divisions de recensement (DR) et des régions métropolitaines de recensement (RMR), Statistique Canada utilise actuellement la méthode des composantes et un modèle de régression. Les estimations par régression pour les années 1982, 1983 et 1985 ont été publiées selon les délais prévus dans le numéro 91-211 au catalogue, mais celles de 1984 ont pu être diffusées seulement en mars 1985 à cause d'un retard dans la collecte des données sur les allocations familiales. En outre, comme on le verra plus loin, ces données présentaient des lacunes sur le plan de la fiabilité. Par conséquent, comme les estimations de la population des RMR fondées sur cette information n'étaient pas acceptables, il a fallu appliquer une autre méthode.

Les estimations de la population des DR et des RMR, calculées par la méthode des composantes, ont été publiées dans le numéro 91-212 au catalogue comme prévu relativement aux années 1982 et 1983; les estimations relatives à 1984 devaient paraître en avril 1986. L'évaluation de ce genre d'estimations produites jusqu'à maintenant révèle que la qualité de ces données est bonne.

2. MODIFICATIONS

Depuis la mise en oeuvre du programme d'estimations par régression de la population des DR et RMR, en 1982, certaines modifications ont dû être apportées aux données et aux méthodes de travail:

- Pour le calcul des estimations de 1983 relatives à la DR de Chicoutimi et à la RMR de Chicoutimi-Jonquière, les données sur les allocations familiales ont été corrigées en fonction du niveau de croissance enregistré l'année précédente. La source du problème était liée aux codes postaux utilisés pour recueillir les données sur les allocations familiales.
- En 1984, les estimations relatives à 17 divisions de recensement ont été imputées en fonction d'estimations provisoires calculées par la méthode des composantes.
- En 1984, il a été décidé de calculer les estimations de la population de la RMR de Calgary à partir des résultats du recensement effectué dans cette ville, et d'étendre ce projet pour toute la période 1981-1986 visée.

¹ Version abrégée du document présenté aux réunions du Comité fédéral-provincial sur la démographie, les 28 et 29 novembre 1985, à Ottawa, Canada.
² Rosemary K. Bender, Division de la démographie, Direction du recensement et de la statistique démographique, Statistique Canada, 4^e étage, Immeuble Jean-Talon, Parc Tunney, Ottawa, Ontario, Canada K1A 0T6.

4. CONCLUSION

Trois séries d'estimations par âge et par sexe ont été produites, pour les DR et les RMR, au moyen de la méthode des composantes par cohorte. Pour chaque série, on utilise une composante différente de la migration par âge et par sexe: i) pour la première, le fichier des données fiscales, ii) pour la deuxième, les données sur la mobilité obtenues à partir du dernier recensement, et iii) pour la troisième, les données sur la migration résiduelle obtenues à partir de deux recensements consécutifs.

Bien que les trois répartitions par âge des migrants diffèrent, les estimations de la population par âge et par sexe qu'ils produisent se ressemblent beaucoup. Chaque série comporte ses propres hypothèses. L'utilisation de la répartition par âge des migrants fondée sur les données sur la migration résiduelle pour produire des estimations postcensitaires suppose que la répartition par âge demeure la même tout au long de la période d'estimation. L'utilisation des données sur la mobilité par âge et par sexe pour les années postcensitaires repose sur la même hypothèse. L'utilisation des données fiscales suppose que la répartition par âge et par sexe demeure la même entre deux années consécutives. Ces séries de données ne mesurent toutefois pas le même type de mouvement. Les données sur la migration résiduelle mesurent seulement le mouvement net entre deux recensements consécutifs (c'est-à-dire, dans ce cas-ci, les recensements de 1976 et de 1981). Les données obtenues à partir des réponses à la question sur la mobilité mesurent également le résultat de mouvements sur une période de cinq ans. Par contre, les fichiers de données fiscales indiquent le mouvement survenu pendant une période d'environ 12 mois. À partir des comparaisons exposées dans le présent document, il est impossible de conclure qu'une série de données sur les migrations produit de meilleures estimations de la population qu'une autre. Une évaluation plus satisfaisante des trois séries d'estimations ne pourra être faite que lorsque les résultats du prochain recensement seront connus.

BIBLIOGRAPHIE

- CENTRAL STATISTICS BUREAU (1980), *Population Estimates by Age Groups for School Districts, 1977-1980*. Rapport technique non publié, gouvernement de la Colombie-Britannique.
- NORRIS, D., et STANDISH, L. (1983). Rapport technique sur la production des données migratoires à partir des dossiers des impôts. Rapport technique, Division de l'exploitation des données administratives, Statistique Canada.
- STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population suivant l'état matrimonial, l'âge, le sexe et composantes de l'accroissement, Canada et provinces*, vol. 2, deuxième édition. N° 91-210 au catalogue, Ottawa: ministère des Approvisionnement et Services.
- STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population des divisions et régions métropolitaines de recensement (méthode de régression)*. N° 91-211 au catalogue, Ottawa: ministère des Approvisionnement et Services.
- STONE, LEROY O. (1980). Evaluating the Relative Accuracy and Significance of Net Migration Estimates. *Demography*, 4, 310-330

Tableau 4

Répartition des régions métropolitaines de recensement selon le niveau de dissémbalance obtenu en comparant les répartitions par âge de la population calculées à partir des trois sources de données: données sur la migration résiduelle, données sur la mobilité et données sur la migration tirées des données fiscales, 1982-1984

Année/	Masculin			Féminin		
Indice de dissémbalance	Migration résiduelle et mobilité	données fiscales	Mobilité et données fiscales	Migration résiduelle et mobilité	données fiscales	Mobilité et données fiscales

ANNÉE 1982	0-3	24	24	24	23	22
	3-5	0	0	0	0	1
	5+	0	0	1	1	24
Total	24	24	24	24	24	24

ANNÉE 1983	0-3	22	23	22	21	20
	3-5	2	0	0	1	2
	5+	0	1	2	2	1
Total	24	24	24	24	24	24

ANNÉE 1984	0-3	21	21	21	20	20
	3-5	2	2	2	0	1
	5+	1	1	1	3	4
Total	24	24	24	24	24	24

Source: Division de la démographie, Statistique Canada, octobre 1985.

(dans ce cas-ci, les estimations emboîtées de la population), produits dans les six mois après la date de référence, on pourrait établir, avec les estimations des naissances, des décès par âge et des migrations nettes fractionnées par âge, de la façon décrite dans les parties 2.1 à 2.4, les estimations de la population par âge et par sexe pour les DR et les RMR, au plus tard huit mois après la date de référence.

3.3 Cohérence

Par cohérence, on entend l'homogénéité des sources d'ensembles de données utilisées pour produire les estimations à divers niveaux de désagrégation administrative ou à divers niveaux de répartition géographique et l'uniformité des méthodes d'estimation. Même s'il faut utiliser dans certains cas une méthode différente, il est de loin préférable d'utiliser toujours la même méthode pour assurer la cohérence méthodologique entre les divers niveaux de désagrégation géographique.

Au niveau provincial ainsi qu'au niveau des DR et des RMR, les sources de données sont les mêmes en ce qui a trait aux naissances, aux décès c'est-à-dire, les statistiques de l'état civil. Au niveau des données sur la migration, les sources sont également les mêmes, à savoir les données fiscales et les données sur la mobilité obtenues à partir du recensement de 1981, pour tous les niveaux de désagrégation géographique. Une autre série de données, soit la série de données sur la répartition par âge des migrants fondée sur la méthode des données sur la migration résiduelle obtenues à partir de deux recensements consécutifs, est toutefois également utilisée.

A tous les niveaux géographiques on utilise comme méthodologie celle des composantes par cohorte.

Tableau 2
Pourcentage des divisions de recensement par province
et des RMR présentant des observations aberrantes^a, 1981

Provinces	Hommes		Femmes	
	R	M	R	M
Terre-Neuve	0	0	0	10
Île-du-Prince-Édouard	0	0	33	0
Nouvelle-Écosse	17	28	11	28
Nouveau-Brunswick	13	7	7	7
Québec	17	16	20	20
Ontario	13	21	13	23
Manitoba	22	22	22	43
Saskatchewan	11	6	11	22
Alberta	13	13	7	7
Colombie-Britannique	14	17	14	14
Total	15	16	15	20
RMR	8	21	67	21

Note: R: Répartition par âge des migrants fondée sur la méthode des résidus.
M: Répartition par âge des migrants fondée sur la mobilité.

F: Répartition par âge des migrants fondée sur les données fiscales.

^a Les observations aberrantes correspondent aux DR qui affichent une EAM de plus de 10% et aux RMR dont l'EAM est supérieur à 5%.

Source: Tableau 1.

Répartition des divisions de recensement selon le niveau de dissémbalance obtenu en comparant les répartitions par âge de la population calculées à partir des trois sources de données: données sur la migration résiduelle, données sur la mobilité et données sur la migration tirées des données fiscales, 1982-1984

Année/ Masculin	Fémnin	
	Migration résiduelle et données fiscales	Mobilité résiduelle et données fiscales

ANNÉE 1982	0-5	245	237	242	240	241	234
	5-10	7	13	10	8	5	11
	10 +	8	10	8	12	14	15
Total		260	260	260	260	260	260

ANNÉE 1983	0-5	235	221	223	230	229	223
	5-10	11	18	21	10	13	13
	10 +	14	21	16	20	18	24
Total		260	260	260	260	260	260

ANNÉE 1984	0-5	240	226	229	235	233	231
	5-10	11	16	14	15	13	12
	10 +	9	18	17	10	14	17
Total		260	260	260	260	260	260

Le tableau 2 montre la répartition en pourcentage des DR et des RMR pour lesquelles on enregistre des observations aberrantes. On considère comme observation aberrante une FAM de plus de 10% dans le cas des DR et de plus de 5% dans le cas des RMR. Les données sont classées par sexe et selon les trois types de répartition par âge des migrants. Comme on s'y attendait, aussi bien pour les hommes que pour les femmes, le pourcentage de cas d'observations aberrantes est en général peu élevé lorsque les estimations sont produites au moyen de la répartition par âge des migrants calculée selon la méthode fondée sur les données sur la migration résiduelle. Par contre, le pourcentage de cas d'observations aberrantes tend à être élevé lorsque les estimations sont produites au moyen de la répartition des migrants fondée sur les données fiscales.

Stabilité temporelle des trois séries d'estimations dans les années postcensitaires, 1982-1984

Comme il n'y a pas de répartitions par âge standard avec lesquelles les estimations de la population peuvent être comparées pendant les années postcensitaires, nous comparons entre elles les trois séries d'estimations de la population par âge et par sexe pour examiner la stabilité temporelle entre elles. Un indice sommaire, l'indice de dissémination, correspond à la moitié de la somme des écarts absolus entre les estimations de la population produites à partir de deux distributions (en pourcentage) par âge des migrants, est utilisé à cette fin. L'indice varie de 0 à 100. Plus la valeur de l'indice est faible, plus les deux distributions comparées sont semblables. Les petites régions sont classées en trois niveaux de dissémination: i) le plus petit niveau de différence, c.-à-d. des indices de 0 à 5%, ii) le niveau moyen de différence, indices de 5 à 10%, et iii) les observations aberrantes, c'est-à-dire les régions dont l'indice de dissémination est égal ou supérieur à 10%. Le classement des DR selon la valeur de l'indice de dissémination figure au tableau 3 et celui des RMR, au tableau 4.

Si on examine le tableau 3, il semble que les trois types de répartition par âge de la population tendent à être semblables et qu'en moyenne, un très grand pourcentage de cas, environ 90%, se situent dans la catégorie des faibles indices de dissémination (0 à 5%), alors que seulement 7% se situent dans la catégorie des indices de 5 à 10%. Nous avons aussi examiné, pour les dix provinces et leur total de population, le pourcentage de cas où le niveau de différence est très élevé (indice de dissémination excédant 10%). Chez les hommes, les pourcentages de cas extrêmes sont faibles, soit de 3 à 5%, entre la répartition par âge calculée selon la méthode fondée sur les données sur la migration résiduelle et la répartition par âge calculée à partir des données sur la mobilité. Chez les femmes, on note une proportion relativement plus élevée d'observations aberrantes. Pour les deux autres comparaisons, soit la comparaison entre les estimations produites selon la méthode fondée sur les données sur la migration résiduelle et les estimations produites selon la méthode fondée sur les données sur la mobilité et les estimations produites selon la méthode fondée sur les données fiscales, on trouve une proportion légèrement plus importante de cas d'observations aberrantes. Les résultats sont semblables pour les régions métropolitaines de recensement (voir tableau 4).

En conclusion, on peut dire que, même si les trois répartitions par âge des migrations (fondées sur les données sur la migration résiduelle, sur les données sur la mobilité et sur les données fiscales) sont différentes, les répartitions par âge de la population qui en résultent se ressemblent beaucoup.

3.2 Délai de production

Par délai de production, on entend la production d'estimations dans le délai le plus court possible après la date de référence. En utilisant les chiffres provisoires de la population

Tableau 1
Répartition des divisions de recensement et des RMR selon la
précision des estimations de la population par âge, 1981

Méthodes d'évaluation de la migration		Niveau d'erreur absolue moyenne (%) par sexe				
		Masculin		Féminin		
		10 +	5-10	3-5	Moins de 3	10 +
Provinces						
Terre-Neuve	R	8	2	0	2	0
	M	2	7	1	3	1
Ile-du-Prince-Édouard	R	1	1	0	0	1
	M	1	2	0	2	0
Nouvelle-Écosse	R	8	4	5	3	2
	M	3	6	6	2	5
Nouveau-Brunswick	R	10	1	4	3	1
	M	4	7	8	3	1
Québec	R	13	27	18	19	15
	M	12	26	23	21	15
Ontario	R	30	8	5	4	7
	M	8	16	10	10	12
Manitoba	R	4	6	6	8	5
	M	1	5	6	7	10
Saskatchewan	R	10	5	5	2	2
	M	1	11	5	5	4
Alberta	R	9	3	3	3	1
	M	5	5	4	5	1
Colombie-Britannique	R	19	3	1	1	4
	M	9	13	8	3	4
RMR	R	14	8	3	2	0
	M	2	17	5	9	1

Note: R: Répartition par âge des migrants fondée sur la méthode des résidus, 1976-1981.
M: Répartition par âge des migrants fondée sur la mobilité, 1981.
F: Répartition par âge des migrants fondée sur les données fiscales annuelles.
Source: Division de la démographie, Statistique Canada, 1985.

par âge des migrants calculée au moyen de la méthode fondée sur les données sur la migration résiduelle, soient comparables aux estimations de la population fondées sur la répartition par âge calculée à partir du recensement de 1981. On ne peut donc pas conclure que la répartition par âge des migrants fondée sur les données sur la migration résiduelle produit de meilleures estimations que la répartition par âge des migrants fondée sur les réponses à la question sur la mobilité ou de meilleures estimations que la répartition par âge des migrants calculée à partir des données fiscales.

3. EVALUATION

L'évaluation est faite en fonction de trois critères: i) la précision, ii) les délais de production et iii) la cohérence. Chacun des critères est examiné en détail dans la partie qui suit.

3.1 Précision

La précision des estimations de la population par âge et par sexe dépend dans une large mesure de la qualité des estimations de la répartition par âge et par sexe des migrants, car les données sur les décès par âge et par sexe sont jugées satisfaisantes. Par conséquent, l'évaluation des estimations de la population par âge et par sexe jette indirectement de la lumière sur l'exactitude des estimations des migrations par âge et par sexe. Pour évaluer l'exactitude des estimations de la population, on compare ces estimations aux chiffres de recensement correspondants.

Pour chaque DR et chaque RMR, les trois séries d'estimations de la population par âge et par sexe, au 1 juin 1981, qui ont été produites à l'aide de la répartition par âge des migrants établie à partir des trois sources, la migration résiduelle (1976-1981), la mobilité (1981) et les données annuelles du fichier sur les impôts et les chiffres des fichiers des allocations familiales, de la façon décrite dans les sections 2.1 à 2.4, ont été comparées aux chiffres du recensement de 1981. Les différences observées ont été considérées comme des erreurs et, pour chaque petite région, on a calculé un indice sommaire appelé "erreur absolue moyenne" (EAM) en prenant la moyenne arithmétique des erreurs en pourcentage (sans tenant compte du signe) pour les seize groupes d'âge de cinq ans. Plus la valeur de cet indice est petite, plus les estimations sont exactes. Au tableau 1, un classement des DR par province et des RMR est présenté pour quatre niveaux d'erreur absolue moyenne: moins de 3%, 3-5%, 5-10% et plus de 10%. Dans l'ensemble, il ressort de ce tableau que la répartition par âge des migrants fondée sur les données sur la migration résiduelle produit de meilleures estimations. Pour les personnes de sexe masculin, environ 66% des divisions de recensement présentent une EAM inférieure à 5%. Dans le cas des personnes de sexe féminin, ce pourcentage est légèrement supérieur, soit 69%. En revanche, on observe des pourcentages inférieurs dans les estimations de la population produites à partir des données sur la mobilité (55 et 57% respectivement pour les hommes et les femmes) et des données sur les migrations obtenues à partir des données fiscales (9 et 19% respectivement pour les hommes et les femmes). Pour les RMR également, la répartition par âge des migrants calculée à partir des données fiscales est inférieure à 3% était de 58% dans le cas des hommes et de 79% dans le cas des femmes. La répartition fondée sur la mobilité et celle qui est fondée sur les données fiscales viennent respectivement au deuxième et au troisième rang. Cela vaut tant pour les hommes que pour les femmes.

Dans toutes les provinces, à l'exception de l'Île-du-Prince-Édouard, la précision des trois séries de répartition par âge des migrants semble bonne. Cette observation vaut tant pour les deux sexes. Cependant, dans certains cas, la répartition par âge des migrants fondée sur les données sur la mobilité semblait donner des résultats semblables. Dans le cas des hommes, on a observé une similitude entre les résultats des deux méthodes dans trois provinces, soit Terre-Neuve, le Nouveau-Brunswick et la Colombie-Britannique, alors que pour les femmes, c'était le cas seulement au Nouveau-Brunswick. Il convient de souligner que, pour la répartition par âge des migrants établie au moyen de la méthode fondée sur les données sur la migration résiduelle, on utilise les répartitions par âge calculées à partir des recensements de 1976 et 1981. Par conséquent, on peut s'attendre que les estimations de la population au 3 juin 1981, produites à partir de la répartition

obtenue à partir des données du dernier recensement. Cette méthode de régression qui produit des estimations emboîtées est décrite dans la publication n° 91-211 au catalogue de Statistique Canada. Pour chaque sexe, la migration nette résiduelle est calculée au moyen de la différence entre les variations démographiques et l'accroissement naturel. Cette migration est ensuite répartie pour chaque région par groupe d'âge de 5 ans au moyen de données sur les migrations par âge provenant de trois sources: migration nette résiduelle selon les recensements de 1976 et 1981, données sur les migrations tirées du fichier de l'impôt sur le revenu et réponses à la question du recensement de 1981 sur la mobilité, "Où habitez-vous il y a 5 ans, c'est-à-dire le 1 juin 1976?". À partir des réponses données à cette question, on peut calculer le nombre de migrants entrants et sortants de chaque petite région. Enfin, la migration par groupe d'âge de 5 ans est répartie en migration par année d'âge au moyen de multiplicateurs de SPRAQUE. Avant d'appliquer ces multiplicateurs, il faut d'abord décomposer la migration nette résiduelle en migration d'entrée et en migration de sortie. En utilisant comme référence les données fiscales pour le calcul des migrations d'entrée et de sortie, on refait ensuite ce calcul pour chaque groupe d'âge de 5 ans.

$$\text{Migration d'entrée résiduelle} = \frac{\text{migration nette selon les données de l'impôt sur le revenu}}{\text{migration nette résiduelle}} \times \text{migration nette résiduelle}$$

Migration de sortie résiduelle = migration d'entrée résiduelle - migration nette résiduelle

Quand les ratios décrits ci-dessus sont utilisés, d'importants problèmes se présentent lorsqu'on la migration nette fractionnée n'a pas le même signe que la migration nette calculée à partir des données fiscales. Dans ce cas, on conserve le signe de la migration nette fractionnée, mais les entrées et les sorties qui en résultent sont échangées de façon à produire le signe approprié. Cette opération est fondée sur l'hypothèse de l'ampleur égale d'un renversement du flux migratoire.

2.3 Chiffres tirés du fichier des bénéficiaires d'allocations familiales âgés de 1 à 14 ans

Les estimations de la population produites au moyen de la méthode des composantes et des cohortes pour les groupes d'âge 1-4, 5-9 et 10-14 ans sont remplacées par les chiffres sur les bénéficiaires d'allocations familiales des mêmes âges qu'il est facile d'obtenir, pour les DR et les RMR, 3 ou 4 mois après la date de référence. Le programme des allocations familiales est universel au Canada et les chiffres obtenus à partir du fichier des allocations familiales sont jugés complets à toutes fins pratiques. Puisque ces données ne sont pas disponibles par sexe, elles sont réparties par sexe selon la répartition par sexe obtenue à partir des données du dernier recensement.

2.4 Ajustements pour assurer la cohérence des estimations provinciales et des estimations par division de recensement

Les estimations postcensitaires emboîtées de la population totale de chaque DR et de chaque RMR sont diffusées dans les 6 mois de la date de référence. De plus, les estimations provinciales sont aussi produites par âge et par sexe dans à peu près le même délai. Les estimations de la population par âge et par sexe obtenues de la façon décrite ci-dessus pour les DR de chaque province sont comparées aux estimations de la population totale des divisions de recensement et aux estimations provinciales par âge et par sexe, et les différences sont distribuées proportionnellement. Pour les régions métropolitaines de recensement, les totaux par âge et par sexe ne sont corrigés qu'en fonction des estimations de la population totale des RMR.

où f_0 = facteur de séparation pour les décès à l'âge 0

$$M_a = \text{migrants nets d'âge } a \text{ entre l'année } t \text{ et l'année } t + 1$$

$$N = \text{naissances entre l'année } t \text{ et l'année } t + 1$$

$$D_a = \text{décès à l'âge } a \text{ entre l'année } t \text{ et l'année } t + 1$$

$$P'_a = \text{population d'âge } a \text{ l'année } t.$$

La méthode des composantes par cohorte (publication n° 91-210 au catalogue) est également utilisée par Statistique Canada pour produire des estimations provinciales et par la Colombie-Britannique pour obtenir des estimations par âge de la population des divisions de recensement, des districts scolaires et des districts de santé (Central Statistics Bureau, 1980).

2.2 Préparation des données d'entrée de base

Puisque nous proposons de produire des estimations postcensitaires provisoires de la population dans les 8 mois de la date de référence, nous ne pouvons utiliser les données définitives sur les composantes de l'accroissement démographique de la population parce que celles-ci ne sont connues qu'après un délai de 18 à 24 mois. En conséquence, il faut utiliser des estimations pour chaque composante.

Naissances et décès

Les estimations provisoires des naissances par sexe pour l'année t sont obtenues en multipliant la répartition provinciale en pourcentage par sexe des naissances dans les petites régions pour l'année $t - 1$ par le total provisoire des naissances dans la province pour l'année t . De la même façon, on obtient les estimations provisoires des décès par âge et par sexe pour l'année t en multipliant la répartition provinciale des décès par âge et par sexe dans les petites régions pour l'année $t - 1$ par le total provisoire des décès dans la province pour l'année t . Enfin, on convertit les décès en décès par cohorte en posant comme hypothèse que les dates de naissance de ceux qui meurent et le nombre de décès sont répartis uniformément sur 12 mois à l'exception des décès à l'âge 0. Les formules correspondantes sont les suivantes:

Pour l'âge 0:

$$\text{Décès par cohorte (0)} = \text{décès (0)} \times 0.89$$

Pour l'âge 1,

$$\text{Décès par cohorte (1)} = [\text{décès (0)} \times 0.11] + [\text{décès (1)} \times 0.5]$$

Pour les âges 2 à 84,

$$\text{Décès par cohorte (âge)} = [\text{décès (âge-1)} \times 0.5] + [\text{décès (âge)} \times 0.5]$$

$$\text{Décès par cohorte (âge 85 et plus)} = \text{décès (84)} \times 0.5 + \text{décès (85 et plus)}.$$

Dans les formules ci-dessus, les facteurs de séparation (f) sont de 0.89 pour l'âge 0, de 0.11 pour l'âge 1 et de 0.5 pour tous les autres âges.

Migration nette résiduelle

Tout d'abord, les estimations postcensitaires de la population totale des DR et des RMR calculées par la méthode de régression sont réparties par sexe selon la répartition par sexe

Estimation de la population par âge et par sexe des divisions de recensement et des régions métropolitaines de recensement¹

RAVI B.P. VERMA, K.G. BASAVARAJAPPA, et ROSEMARY K. BENDER²

RÉSUMÉ

Une méthode d'estimation de la population par année d'âge et par sexe a été élaborée pour les petites régions (divisions de recensement et régions métropolitaines de recensement). Pour améliorer leur fiabilité, les estimations démographiques par année d'âge ont été classées par groupe de cinq années d'âge et seules ces données peuvent être diffusées. Elles sont fondées sur la composition par âge et par sexe de la population selon le dernier recensement, les naissances par sexe, les décès par année d'âge et par sexe, les estimations des migrations par âge et par sexe et les bénéficiaires d'allocations familiales âgés de 1 à 14 ans.

MOTS CLÉS: Méthode des composantes par cohorte; erreur absolue moyenne; indice de dissemblance; facteur de séparation.

1. INTRODUCTION

Le présent document décrit la méthode utilisée pour produire des estimations de la population par âge et par sexe des petites régions (divisions de recensement et régions métropolitaines de recensement), donne les résultats de l'examen des méthodes d'estimation et, enfin, traite des facteurs qui influent sur la qualité des estimations. Selon les chiffres du recensement de 1981, la population des 266 divisions de recensement variait de 2,000 à 2,000,000 d'habitants et celle des 24 régions métropolitaines de recensement, de 100,000 à 3,000,000 d'habitants. La section 2 décrit les méthodes d'estimation et les principales sources de données. La section 3 donne les résultats de l'évaluation des estimations relatives à la migration et à la population.

2. MÉTHODES

Dans cette section, nous décrivons les méthodes d'estimation et la préparation des données d'entrée de base.

2.1 Méthode des composantes par cohorte

Pour chaque division de recensement (DR) et chaque région métropolitaine de recensement (RM), nous utilisons la méthode des composantes par cohorte pour estimer la population par âge. Les équations correspondantes sont les suivantes:

- (1) Pour l'âge 0, $P_{0+1}^0 = N - f_0 D_0 + \frac{1}{2} M_0$
- (2) Pour l'âge 1, $P_{1+1}^1 = P_0^1 - [(1 - f_0) D_0 + \frac{1}{2} D_1] + \frac{1}{2} (M_0 + M_1)$
- (3) Pour les âges 2 à 84, $P_{a+1}^{a+1} = P_a^a - \frac{1}{2} (D_a + D_{a+1}) + \frac{1}{2} (M_a + M_{a+1})$
- (4) Pour les âges 85 et plus, $P_{85+}^{85+} = P_{84+}^{84+} - \frac{1}{2} D_{84} - D_{85+} + \frac{1}{2} M_{84} + M_{85+}$

¹ Version abrégée du document présentée aux réunions du Comité fédéral-provincial sur la démographie tenues les 28 et 29 novembre 1985 à Statistique Canada, Ottawa, Canada. Cette recherche a été faite en collaboration avec le personnel du Programme de données régionales de Statistique Canada.
² Ravi B.P. Verma, K.G. Basavarajappa et Rosemary K. Bender, Division de la démographie, Statistique Canada, 4^e étage, Immeuble Jean Talon, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

BIBLIOGRAPHIE

- BUREAU CENTRAL DE LA STATISTIQUE (1982). British Columbia Municipal Population Estimation Methodology. Document non publié. Ministère du Développement de l'industrie et de la petite entreprise, Victoria.
- SPSS INC. (1984). *Statistical Package for the Social Sciences/PC*. Chicago, B265-B280.
- SHRYOCK, H.S., et SIEGAL, J.S. (1980). *The Methods and Materials of Demography*. 2, U.S. Bureau of the Census, 628-630.

6. CONCLUSIONS

La méthode décrite dans ce document présente des avantages dans le cas des régions pour lesquelles il existe de bonnes sources de données chronologiques sur la population et de statistiques de l'état civil. On estime que l'application d'une méthode comportant des estimations de la migration nette est relativement simple, qu'elle produit des taux d'erreur acceptables et permet la production d'estimations par âge et sexe rapidement après la date de référence. Certes, il serait idéal d'établir des estimations de la migration interne et externe, mais on ne dispose pas actuellement d'assez d'information sur les mouvements migratoires dans les petites régions de la Colombie-Britannique. Une autre amélioration qui est envisagée a trait à l'utilisation des chiffres relatifs à la sécurité de la vieillesse pour accroître la stabilité et la précision des estimations concernant les groupes d'âge supérieurs.

REMERCIEMENTS

Les auteurs tiennent à remercier Don McKrae, Steve Miller, Ravi Verma, Garnett Picot et Paul Knapp ainsi que tous leurs prédécesseurs pour leur apport et leur soutien à la construction du système de ventilation des estimations.

Tableau 7
Pourcentage d'erreurs entre les estimations et les chiffres du recensement de 1981, selon la division de recensement, pour tous les groupes d'âge

Division du recensement	Population totale			Total			Hommes			Femmes		
	EAMP	IMD	(%)	EAMP	IMD	(%)	EAMP	IMD	(%)	EAMP	IMD	(%)
1000 East Kootenay	53,725	4.24	2.04	5.24	2.29	3.88	2.15	3.88	2.29	2.15	3.88	2.15
3000 Central Kootenay	52,045	4.00	2.18	4.03	2.13	5.06	1.69	5.06	2.13	1.69	5.06	1.69
5000 Kootenay-Boundary	33,235	2.32	1.23	2.34	1.18	3.21	1.68	3.21	1.18	1.68	3.21	1.68
7000 Okanagan-Similkamien	57,185	5.04	2.64	6.02	3.08	4.72	2.49	4.72	3.08	2.49	4.72	2.49
9000 Fraser-Cheem	56,930	3.12	1.60	3.33	1.78	4.15	2.08	4.15	1.78	2.08	4.15	2.08
1000 Central Fraser Valley	115,015	3.14	1.43	3.46	1.52	3.65	1.81	3.65	1.52	1.81	3.65	1.81
13000 Dowdney-Alouette	62,000	2.10	1.15	2.56	1.23	2.26	1.32	2.26	1.23	1.32	2.26	1.32
15000 Greater Vancouver	1,168,700	1.63	0.94	1.68	0.93	1.67	0.98	1.67	0.93	1.67	0.98	0.98
17000 Capital	249,475	1.64	0.87	2.31	1.21	1.18	0.61	1.18	1.21	0.61	1.18	0.61
19000 Cowichen Valley	45,315	3.09	1.66	3.36	1.69	3.85	2.08	3.85	1.69	2.08	3.85	2.08
21000 Nanaimo	84,815	3.07	1.58	3.40	1.74	3.22	1.66	3.22	1.74	1.66	3.22	1.66
23000 Alberni-Clayoquot	32,560	2.75	1.36	2.88	1.27	3.27	1.68	3.27	1.27	1.68	3.27	1.68
25000 Comox-Strathcona	68,620	1.44	0.80	1.85	0.87	2.85	1.50	2.85	0.87	1.50	2.85	1.50
27000 Powell River	19,050	5.36	2.58	5.06	2.44	6.18	3.03	6.18	2.44	3.03	6.18	3.03
29000 Sunshine Coast	16,625	4.84	2.57	6.79	3.58	5.65	2.81	5.65	3.58	2.81	5.65	2.81
31000 Squamish-Lillooet	18,925	1.82	0.99	2.56	1.37	3.10	1.58	3.10	1.37	1.58	3.10	1.58
33000 Thompson-Nicola	102,430	2.13	1.10	2.07	0.10	2.65	1.37	2.65	0.10	1.37	2.65	1.37
35000 Central Okanagan	85,235	3.96	1.93	3.91	1.88	4.32	2.14	4.32	1.88	2.14	4.32	2.14
37000 North Okanagan	69,033	5.26	2.52	6.44	3.06	5.05	2.50	5.05	3.06	2.50	5.05	2.50
39000 Columbia-Shuswap	45,425	3.04	1.63	3.56	1.84	2.99	1.66	2.99	1.84	1.66	2.99	1.66
41000 Cariboo	58,810	3.18	1.93	3.90	2.18	3.42	2.06	3.42	2.18	2.06	3.42	2.06
43000 Mount Waddington	14,675	8.96	3.04	5.13	1.59	17.77	5.49	17.77	1.59	5.49	17.77	5.49
45000 Central Coast	3,050	17.99	7.62	21.62	8.86	14.92	7.34	14.92	8.86	7.34	14.92	7.34
47000 Skeena-Queen Charlotte	24,030	4.82	2.09	5.70	2.58	4.61	1.84	4.61	2.58	1.84	4.61	1.84
49000 Kitimat-Stikina	41,790	6.26	1.99	4.99	1.66	8.59	2.78	8.59	1.66	2.78	8.59	2.78
51000 Bulkley-Nechako	38,310	6.23	2.31	5.76	2.10	6.83	2.57	6.83	2.10	2.57	6.83	2.57
53000 Fraser-Fort George	89,430	3.50	1.41	3.39	1.25	3.72	1.68	3.72	1.25	1.68	3.72	1.68
55000 Peace River-Liard	55,340	8.00	2.95	9.43	3.65	7.34	2.83	7.34	3.65	2.83	7.34	2.83
57000 Stikine	2,685	17.15	6.89	17.89	6.88	22.39	8.35	22.39	6.88	8.35	22.39	8.35
Erreur moyenne		4.83	2.17	5.19	2.31	5.60	2.51	5.60	2.31	2.51	5.60	2.51

Le tableau 4 (pourcentage d'erreurs dans les divisions de recensement selon la taille de la population) indique qu'à la suite de l'aggrégation au niveau de régions intraprovinciales plus grandes, on obtient une amélioration des taux d'erreur. Le tableau 7 montre qu'il existe une relation négative entre les taux d'erreur et la taille de la population, à l'échelon de la division de recensement.

La comparaison des tableaux 5 et 6 confirme l'amélioration des taux d'erreurs lorsqu'on agrège en fonction de cellules par âge et sexe plus grandes. Bien qu'il faille être prudent dans l'utilisation d'estimations régionales par âge et sexe, les auteurs estiment qu'elles sont fiables.

Tableau 4

Pourcentage d'erreurs pour toutes les divisions de recensement selon la taille de la population

Taille de population	Total			Hommes			Femmes		
	EAMP (%)	IMD (%)		EAMP (%)	IMD (%)		EAMP (%)	IMD (%)	N
0-39,000	7.22	1.94	7.55	2.13	8.79	2.29	10	10	10
40,000-59,999	4.32	1.82	5.03	2.14	4.91	1.83	10	9	9
60,000 +	2.51	0.87	2.73	0.98	2.84	0.90	9	29	29
Moyenne des divisions de recensement	4.83	1.27	5.19	1.41	5.60	1.35			

Tableau 5

Districts scolaires
Nombre d'estimations par intervalle d'erreur

N° de cellules	Pourcentage	Intervalle d'erreur absolue moyenne en pourcentage			Intervalle d'erreur absolue moyenne en pourcentage		
		< 5	5 à 10	10 à 15	15 +	Total	
674	61%	239	101	96	1110	100%	
22%	9%	9%	22%	101	96	1110	

Tableau 6

Divisions de recensement
Nombre d'estimations par intervalle d'erreur

N° de cellules	Pourcentage	Intervalle d'erreur absolue moyenne en pourcentage			Intervalle d'erreur absolue moyenne en pourcentage		
		< 5	5 à 10	10 à 15	15 +	Total	
306	70%	77	25	27	435	100%	
18%	6%	6%	18%	77	25	435	

Comme l'indique le tableau 2, les régions peu peuplées ont en général un pourcentage d'erreur plus élevé. Le pourcentage élevé d'erreurs dans les petites régions peut être attribué à l'instabilité des économies de ressources, laquelle se reflète dans la répartition de la migration nette.

Au niveau des divisions de recensement, on constate qu'il se dégage une tendance semblable. Comme le montre le tableau 3, l'erreur absolue moyenne en pourcentage pour toutes les régions et tous les groupes d'âge est de 4,83% (5,19% pour les hommes et 5,60% pour les femmes). L'IMD est de 1,27% au total, de 1,41% pour les hommes et de 1,35% pour les femmes. L'erreur est encore une fois bimodale, atteignant des sommets dans les groupes des 20-29 ans et des 60-69 ans. En outre, les femmes affichent des taux d'erreur plus élevés que les hommes dans les groupes des 20-29 ans, et des taux d'erreur plus bas dans les groupes des 60-69 ans.

Tableau 2
Pourcentage d'erreurs pour tous les districts scolaires
selon la taille de la population

Taille de population	Hommes		Femmes		Total
	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)	
0-9,999	8.87	3.16	10.14	3.89	10.27
10,000-24,999	6.07	2.47	6.92	2.96	6.62
25,000 +	3.66	1.67	3.92	1.78	4.09
Moyenne pour les districts scolaires	6.20	1.95	7.00	2.15	7.00
					2.08

Tableau 3
Pourcentage d'erreurs entre les estimations et les chiffres du recensement de 1981, selon le groupe d'âge, pour toutes les divisions de recensement

Groupe d'âge	Total		Hommes		Femmes
	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)	EAMP (%)
0-4	2.37	0.54	3.20	0.76	2.28
5-9	1.52	0.50	1.71	0.55	2.13
10-14	1.69	0.39	2.75	0.57	2.50
15-19	3.81	1.39	3.79	1.30	4.68
20-24	9.83	3.07	9.30	2.91	10.90
25-29	7.02	3.04	7.30	2.87	8.09
30-34	3.28	1.29	3.31	1.43	3.85
35-39	3.34	0.66	3.06	0.57	4.21
40-44	3.86	0.88	4.29	1.01	4.16
45-49	2.91	0.70	3.20	0.75	3.75
50-54	4.82	0.64	4.41	0.75	6.10
55-59	5.49	1.34	5.36	1.55	6.94
60-64	7.88	1.95	8.37	2.29	7.94
65-69	8.48	1.89	10.30	2.67	9.79
70 +	6.16	0.81	7.46	1.20	6.73
Moyenne	4.83	1.27	5.19	1.41	5.60
					1.35

où P_{Ei} est la population estimée pour le groupe d'âge i , P_{Ai} est la population d'après le recensement pour le groupe d'âge i et N est le nombre de cellules. L'IMD est défini comme suit:

$$IMD = 100 \times \frac{1}{2} \left[\sum_{i=1}^N (P_{Ai} - P_{Ei}) \right] / \sum_{i=1}^N P_{Ai}$$

où P_{Ai} est la population réelle pour le groupe d'âge i et P_{Ei} est la population estimée pour le groupe d'âge i .

Comme l'indique le tableau 1, l'erreur absolue moyenne en pourcentage pour tous les groupes d'âge et toutes les régions est de 6.20% et l'IMD, de 1.95%. L'erreur moyenne en pourcentage pour les hommes est très semblable à celle pour les femmes (EAMP de 7% dans les deux cas et IMD de 2.15% pour les hommes et de 2.08% pour les femmes). Du point de vue de l'âge, les taux d'erreur les plus élevés figurent dans les groupes des 20-29 ans et des 60-69 ans. Il faut également noter que la répartition par âge des erreurs est différente chez les hommes et chez les femmes. Chez les hommes, les taux d'erreur les plus élevés semblent se trouver dans les groupes d'âge supérieurs, tandis que chez les femmes ils sont dans les groupes très mobiles des 20-29 ans.

Tableau 1

Pourcentage d'erreurs entre les estimations et les chiffres du recensement de 1981, selon le groupe d'âge, pour tous les districts scolaires

Âge	Total		Hommes		Femmes	
	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)
0-4	3.33	.96	3.94	1.21	3.62	1.04
5-9	2.80	.76	3.28	0.88	3.62	1.02
10-14	2.33	.64	3.54	0.84	2.88	0.87
15-19	5.20	2.01	5.68	2.01	6.18	2.24
20-24	13.32	4.77	13.50	4.62	14.54	5.12
25-29	8.31	4.07	8.42	3.70	9.41	4.65
30-34	5.02	2.12	5.42	2.45	5.72	2.06
35-39	4.88	1.33	5.73	1.62	5.38	1.34
40-44	4.52	1.33	5.84	1.51	4.67	1.52
45-49	3.60	1.22	4.47	1.37	4.78	1.49
50-54	5.66	1.33	5.86	1.48	6.68	1.54
55-59	6.11	1.72	6.19	1.78	7.82	1.97
60-64	8.86	2.44	10.35	2.95	8.91	2.17
65-69	10.60	2.66	12.53	3.52	11.44	2.30
70 +	8.49	1.95	10.19	2.35	9.33	1.94
Moyenne	6.20	1.95	7.00	2.15	7.00	2.08

Il n'est pas toujours possible de calculer toutes les cinq répartitions. Ainsi, on ne peut faire si une petite région n'a jamais eu de migration nette négative dans le passé mais si elle en a une pour l'année visée. Dans un tel cas, on se sert uniquement des répartitions que l'on peut calculer.

Des essais basés sur les données du recensement de 1981 ont montré que, des cinq répartitions de la migration nette décrites plus haut, la première (la répartition chronologique par petite région) a l'erreur absolue moyenne en pourcentage la plus faible pour tous les districts scolaires et tous les groupes d'âge; elle est suivie de la deuxième (répartition chronologique par groupe), puis de la troisième, etc. Cependant, bien que la première répartition ait l'EAMP la plus faible, elle n'a pas produit le taux d'erreur le plus bas dans chaque cas. Aussi, une technique de sélection a été appliquée pour remplacer la répartition de la population produite au numéro 1 par celle produite aux numéros 2, 3, 4 ou 5 uniquement dans les cas où la répartition de la population produite au numéro 1 était considérée comme n'étant pas fidèle à la répartition de la population pour l'année choisie. La technique de sélection suivante a été conçue à partir des résultats des tests basés sur le recensement de 1981.

Une fois qu'on a calculé toutes les répartitions possibles, on ajoute chacune à la population de base naturelle et on obtient jusqu'à cinq estimations possibles de la population d'une petite région, selon le sexe et l'année d'âge, pour la période suivante. On examine alors ces estimations de la population par âge et par sexe afin de déterminer celle qui produit le moins de changements dans la structure par âge de la petite région par rapport à l'année précédente. Pour ce faire, on calcule d'abord la différence moyenne en pourcentage entre les structures par âge de chacune des cinq populations possibles au temps $t + 1$ et celles de la population au temps t . On calcule ensuite l'écart-type des moyennes obtenues et on indique au moyen d'un code la répartition qui affiche l'écart-type le plus bas. Si l'écart-type produit à partir de la répartition chronologique par petite région est beaucoup plus grand que l'écart-type minimum (c.-à-d. la répartition signalée), cette répartition chronologique est rejetée. On applique le même procédé à la répartition chronologique par groupe, et ainsi de suite jusqu'à ce qu'une des cinq populations possibles soit retenue.

Une fois la population optimale au temps $t + 1$ calculée pour toutes les petites régions, il reste à faire deux derniers rajustements. En premier lieu, on a remplacé les chiffres relatifs aux groupes d'âge 0-14 ans par ceux provenant des allocations familiales, puis ajusté proportionnellement les populations des autres groupes d'âge afin que la population totale de chaque petite région reste constante. En second lieu, on a ajusté proportionnellement la population pour s'assurer que la répartition par âge de la somme de toutes les estimations relatives aux populations des petites régions corresponde à la répartition par âge établie par Statistique Canada pour la Colombie-Britannique.

5. ÉVALUATION DE LA MÉTHODE

Les tableaux présentés plus loin indiquent les taux d'erreur relatifs aux estimations de la population au 1^{er} juin 1981 et répartis en groupes d'âge quinquennaux de 0 à 70 ans et plus, pour 74 districts scolaires de la Colombie-Britannique et 29 divisions de recensement. On a obtenu les estimations de la population par âge et par sexe pour les divisions de recensement en regroupant les estimations de la population pour les districts scolaires. Pour établir l'exactitude des estimations de la population des petites régions par âge et par sexe obtenues par la méthode décrite plus haut, on a produit des estimations de la population de 1981 par sexe et par groupe d'âge quinquennal de 0 à 70 ans et plus pour 74 districts scolaires, puis on a comparé les résultats obtenus avec les chiffres du recensement de 1981. On a eu recours à deux mesures sommaires pour évaluer l'efficacité des estimations de la population par âge et par sexe. Il s'agit de l'erreur absolue moyenne en pourcentage (EAMP) et l'indice de mauvaise distribution (IMD). L'EAMP est défini comme suit:

$$EAMP = 100 \times \left[\sum_{i=1}^N \left| (P_{Ei} - P_{Ai}) / P_{Ai} \right| \right] / N$$

3. RÉPARTITION CHRONOLOGIQUES DE LA MIGRATION NETTE

Pour chacun des 74 districts scolaires de la Colombie-Britannique, les estimations de la migration nette par âge et par sexe ont été calculées, par la méthode des résidus, pour les périodes de 1961-1966, 1966-1971 et 1971-1976. C'est ce qu'on appelle les répartitions chronologiques par petite région.

Si l'on examine ces répartitions de la migration nette par petite région, on découvre qu'elles sont extrêmement instables dans le temps. Pour réduire les effets de cette instabilité, un certain nombre de mesures ont été prises.

Premièrement, les répartitions de la migration par petite région ont été classées selon qu'elles ont eu lieu au cours d'une période de migration totale nette positive ou d'une période de migration totale nette négative. On a découvert que les répartitions par âge de la migration résiduelle pour de nombreuses régions diffèrent selon que la migration est positive ou négative. On a ensuite regroupé les petites régions présentant des répartitions de migration semblables, pour ensuite calculer les répartitions de la migration nette positive et de la migration nette négative. Elles ont été désignées répartitions chronologiques par groupe. L'analyse de certains groupes d'âge (selon la méthode SPSS/PC) a permis de regrouper les répartitions chronologiques de la migration des petites régions. Cette méthode a permis de grouper la plupart des régions en trois grappes, et huit groupes étaient formés uniquement d'une région chacun. Une fois les régions ainsi regroupées, on a déterminé les répartitions de la migration positive et les répartitions de la migration négative pour les périodes de migration positive et négative les plus récentes.

4. ESTIMATIONS DE LA POPULATION DES PETITES RÉGIONS SELON LE SEXE ET L'ANNÉE D'ÂGE

Comme il a été mentionné dans la section 3, les répartitions chronologiques de la migration nette calculées par la méthode des résidus varient considérablement pour certaines régions. Il semble que cela soit imputable à deux facteurs. Premièrement, de nombreuses régions à l'étude ont des économies de ressources déficientes qui affichent de grandes fluctuations, ce qui entraîne des mouvements des taux de migration. Deuxièmement, le calcul d'une distribution en percentiles d'un paramètre tel que la migration nette ayant des valeurs positives, négatives ou nulles introduit un certain degré d'instabilité.

Pour prévenir la construction d'une répartition chronologique de la migration nette qui ne soit pas représentative de la situation courante dans l'année d'estimation, on a calculé cinq répartitions chronologiques, par sexe, qui ont été ventilées par année d'âge. Une description de ces répartitions est donnée ci-après.

- 1) La première répartition de la migration qui a été choisie est la répartition chronologique par petite région de chaque petite région dont le signe est le même que la migration nette vers cette région.
- 2) La deuxième répartition est la répartition chronologique par groupe du groupe auquel la petite région appartient et dont le signe est le même que celui de la migration nette vers cette région.
- 3) On a obtenu la troisième répartition en additionnant *séparément* la migration, pour la période la plus récente, de toutes les petites régions possédant une migration nette positive et une migration nette négative, puis en calculant la répartition par âge.
- 4) La quatrième répartition est celle de la population de base naturelle pour chaque petite région.
- 5) La cinquième et dernière répartition est la répartition par âge des migrants vers la Colombie-Britannique en général. Comme la migration vers la Colombie-Britannique a été positive au cours de toutes les années visées, cette répartition est positive. On l'a néanmoins utilisée, peu importe que la migration vers une petite région ait été positive ou négative.

Estimation de la répartition par âge et par sexe de la population totale des petites régions¹

DAVID S. O'NEIL et CHRIS D. MCINTOSH²

RÉSUMÉ

Ce document décrit une méthode de production d'estimations par âge et par sexe de l'état de la population des petites régions, à partir d'estimations de la population totale, de données sur les naissances et les décès et d'estimations chronologiques de la migration nette résiduelle. On y présente également une évaluation fondée sur les chiffres du recensement de 1981 concernant les divisions de recensement et les districts scolaires de la Colombie-Britannique.

MOTS CLÉS: Estimations de la population par âge et par sexe; petites régions; migration nette résiduelle.

1. INTRODUCTION

Le Bureau central de la statistique produit actuellement des estimations postcensitaires de la population pour diverses régions infraprovinciales à l'aide d'une méthode de régression (Bureau central de la statistique, 1982). En plus des estimations de la population totale par petite région, il produit des estimations par âge et par sexe.

Ce document expose la méthode permettant à partir de l'estimation de la population totale, d'obtenir des estimations par âge et par sexe pour les régions infraprovinciales de la Colombie-Britannique.

2. VUE D'ENSEMBLE

La méthode retenue pour déterminer la population des petites régions par sexe et par année d'âge comporte deux volets.

Le premier volet consiste à examiner les données chronologiques sur la migration nette résiduelle à partir des chiffres de recensement afin d'établir un certain nombre de répartition de la migration par sexe et par année d'âge pour chaque petite région (Shryock et Siegal 1980).

Le second volet de la méthode consiste à faire vieillir la population de base pour chaque sexe, à ajouter les naissances et à retrancher les décès, ce qui donne pour chaque région une nouvelle répartition de la population que l'on désigne la population de base naturelle. En retranchant la population de base naturelle de la population totale estimée selon le sexe, on obtient un résidu qui est égal à la migration nette selon le sexe si la population dénombrée et les données de l'état civil sont exactes pour les deux périodes. Ce résidu par sexe est distribué par année d'âge au moyen d'une répartition chronologique, puis il est ajouté à la population de base naturelle, ce qui donne l'estimation de la population par âge et par sexe d'une région pour la période suivante.

À cause des courts délais de production des données d'entrée, les estimations de la population totale peuvent être calculées quatre mois après la date de référence (1^{er} juin) et les répartition par âge et par sexe, un ou deux mois plus tard.

¹ Version abrégée du document présentée à la réunion du Comité fédéral-provincial sur la démographie, Ottawa, novembre 28-29, 1985.

² D.S. O'Neil, SRL Sociométrics Resources Ltd., et C.M. McIntosh, InterSoft Resources Ltd., Bureau central de la statistique, ministre du développement de l'industrie et de la petite entreprise de la Colombie-Britannique, 2^e étage, 1405, Douglas Street, Victoria, Colombie-Britannique, Canada, V8W 3C1.

Les opinions exprimées dans ce document sont celles des auteurs et ne représentent pas nécessairement le point de vue du gouvernement de la Colombie-Britannique.

- b) Les données sont obtenues de chaque fournisseur sous une forme déjà agrégée par municipalité. Le principal avantage est que les modifications des limites municipales, qui se produisent régulièrement, paraissent dans les données sans entraîner de travail supplémentaire pour le Bureau.
- c) La majorité des données peuvent être obtenues sous forme ordi-nologique et accompagnées du code postal. Cela permet de les convertir facilement en fonction de régions géographiques autres que les municipalités lorsqu'un triage est effectué par le Fichier principal de conversion des codes postaux du Bureau.
- d) Les données peuvent être obtenues gratuitement de chacun des fournisseurs dans un délai relativement court (2 à 3 semaines).

Inconvénients

- a) Les écarts dans les taux de logements vacants entre l'année de base et l'année d'estimation produisent une distorsion des estimations.
- b) Les immeubles d'habitation qui passent d'un simple compteur à des compteurs multiples parfois entre l'année de base et l'année d'estimation produisent une surestimation.
- c) Les régions en mutation, c'est-à-dire qui passent par exemple d'une population saisonnière à une population "stable", introduisent un biais dans les estimations.
- d) Les données proviennent de sources externes et différentes, ce qui pourrait causer éventuellement des problèmes sur le plan de la qualité et de la comparabilité des données, et produire également une situation où l'ordre de priorité du programme des estimations de la population du Bureau serait subordonné aux besoins administratifs d'un organisme externe.

BIBLIOGRAPHIE

BUREAU CENTRAL DE LA STATISTIQUE (1982). British Columbia municipal population estimation methodology. Document non publié, Bureau central de la statistique, ministère du Développement de l'industrie et de la petite entreprise, gouvernement de la Colombie-Britannique.

McRAE, D. (1985). A regression approach to small area population estimation. Communication présentée au colloque international sur les statistiques régionales, Ottawa, Canada, 22-24 mai 1985.

McRAE, D. (1982). British Columbia small area estimation model - 1981 municipal and census division evaluation. Document non publié, Bureau central de la statistique, ministère du Développement de l'industrie et de la petite entreprise, gouvernement de la Colombie-Britannique.

explosion de la construction lorsque les promoteurs ont construit des logements en prévision d'un afflux de population. Chaque logement, occupé ou non, ou même en construction, devait être muni d'un compteur qui était peut-être peu utilisé, mais qui n'en demeurerait pas moins actif, et par là fait même compris dans les calculs. Par conséquent, la variation de la proportion de compteurs par rapport à la population entre 1976 et 1981 a été surévaluée, ce qui a entraîné une surestimation de la population de 1981 pour de nombreuses localités de Peace River-Liard.

Un autre inconvénient de l'utilisation des données de l'Hydro découle de la possibilité d'une conversion, dans le cas des immeubles d'habitation, d'un compteur unique en un réseau de compteurs multiples. Cette situation peut se présenter lorsqu'un ancien immeuble d'appartements, par exemple, qui possédait un seul compteur, est rénové ou transformé en logements à compteurs individuels. Ce changement entraînerait une surestimation de la population dans un modèle de régression, s'il devait survenir à un moment quelconque entre l'année de base et l'année d'estimation.

Enfin, certains problèmes apparaîtront si on utilise les données de l'Hydro pour des régions où la population est en mutation, c'est-à-dire où le rapport entre la population et les compteurs résidentiels se modifie. On peut citer comme exemple en Colombie-Britannique le lieu de villégiature de Whistler. Il y a quinze ans cette municipalité était constituée essentiellement de chalets d'hiver construits aux abords d'une pente de ski. Au cours de la dernière décennie, cependant, elle s'est transformée en un lieu résidentiel à long terme. Par conséquent, le nombre de personnes par compteur de l'Hydro, qui à l'origine était très faible par rapport à la moyenne de la Colombie-Britannique, se rapproche maintenant de la norme. Comme pour le cas des logements vacants, l'utilisation des données de l'Hydro pour estimer la population dans un territoire de ce genre donnerait probablement lieu à des erreurs supérieures à la moyenne. La solution aux trois problèmes mentionnés ci-haut est la suppression des comptes correspondants à une utilisation mensuelle ou bimensuelle faible qui par le fait même représentent des logements considérés vacants. Le Bureau central de la statistique étudie actuellement la possibilité d'appliquer cette mesure dans le cas des données provenant de la B.C. Hydro. Si c'est possible, nous espérons disposer de l'ensemble de données améliorées pour la comparaison avec les résultats du recensement de 1986. Pour l'instant, à titre de solution partielle, les données de l'Hydro pour les secteurs qui en 1981 affichaient un rapport faible ou élevé de personnes par compteur comparativement à la norme provinciale (c.-à-d. moins de 2 ou plus de 5) ne sont pas utilisées.

Une dernière lacune possible est le recours à des sources d'information externes et différentes. Dans le passé, cette situation n'a généralement pas posé de problème. Toutefois, il y a eu quelques rares cas qui ont suscité une remise en question de la qualité des données des compteurs recueillies sur le terrain. Il pourrait s'agir d'une modification des limites d'une municipalité non reflétée dans les données relatives aux compteurs ou de l'adjonction aux données d'un certain type de comptes non résidentiels (par exemple pour les lampadaires). Il importe donc de vérifier soigneusement les données.

6. CONCLUSIONS

Les avantages et les inconvénients de l'utilisation des données de l'Hydro dans le modèle d'estimation par régression de la population en Colombie-Britannique sont les suivants:

Avantages

- a) Les données de l'Hydro, lorsqu'elles sont utilisées dans un modèle de régression, produisent une erreur absolue moyenne en pourcentage plus faible que les données sur les allocations familiales, pour les petites régions.

Tableau 1
Comparaison des erreurs d'estimation entre les sources de données pour les municipalités de la Colombie-Britannique - 1981

Population		Source de données			
		EAMP ^a	Ensemble	n	≥ 4000
		n	EAMP	n	< 4000
		n	EAMP	n	< 4000
MCD/H/F	70	5.53	2.99	88	8.72
MCD/H	70	5.16	4.04	88	6.58
MCD/F	75	10.46	4.57	92	17.69

$$EAMP = \left[\frac{\sum_{i=1}^n \left| \frac{Y_i}{Y_i - Y_i} \right| \right] \div n \times 100$$

ou:

Y_i = population du recensement pour la région i
 Y_i = population estimée pour la région i
 n = nombre de régions estimées.

Tableau 2

Test pour déterminer les différences statistiquement significatives entre les erreurs moyennes en pourcentage, selon certaines sources de données - 1981

Intervalle de confiance de 95% pour la différence moyenne dans les erreurs absolues en pourcentage		Source de données			
		Ensemble	≥ 4000	< 4000	Population
MCD/H/F - MCD/H	.37 ± .86	-1.05 ± .56 ^a	2.14 ± 1.76 ^a		
MCD/H/F - MCD/F ^b	-4.86 ± 1.55 ^a	-1.57 ± .85 ^a	-9.00 ± 3.11 ^a		

^a Différences statistiquement significatives au niveau de 5% calculées à l'aide d'un test-t bilatéral et d'échantillons appariés, en supposant une distribution normale des moyennes.
^b Pour appairer les échantillons, sur 167 estimations possibles fondées sur les comptes d'allocations familiales, 158 seulement ont été utilisées. Le nombre des observations est le suivant: ensemble, 158; plus grand que ou égal à 4,000, 88; moins de 4,000, 70.

L'estimation à la fois dans les grandes et les petites régions. Les données des comptes d'allocation familiales, par contre, amélioreraient la précision pour les régions de grande taille, mais la réduit pour les régions de petite taille, sans produire un effet global statistiquement significatif.

5. INCONVÉNIENTS DES DONNÉES DE L'HYDRO DANS UN MODÈLE DE RÉGRESSION

Un des problèmes qui se posent lorsqu'on utilise les données de l'Hydro dans un modèle d'estimation de la population par régression concerne les logements vacants, ou plus précisément les écarts importants des taux de logements vacants entre l'année de base et l'année d'estimation. Cette lacune a été démontrée dans l'évaluation de 1981 portant sur les localités de la région Peace River-Liard en Colombie-Britannique. Par suite du projet de North-East Coal, les localités de la division de recensement Peace River-Liard en 1981 ont connu une

3. PRÉSENTATION DES DONNÉES

Les principaux fournisseurs d'électricité sont la B.C. Hydro et la West Kootenay Power and Light. Les autres administrations achètent l'électricité aux deux grands fournisseurs et la revendent à leurs propres clients (normalement les résidents de la municipalité).

Sur les neuf sources de données des comptes résidentiels de l'Hydro, seulement celles qui proviennent de la B.C. Hydro sont ordinolinguos, c'est-à-dire lisibles par machine. Les huit autres administrations fournissent les données totalisées par municipalité (urbaine), et un total pour les clients ruraux (non municipaux) s'il y a lieu. La date de référence pour toutes les données est la date de facturation du 31 mai. Dans la plupart des cas, les données peuvent être obtenues dans les 2 à 3 semaines qui suivent la date de facturation.

Les données fournies par la B.C. Hydro se présentent sous deux formes. La première indique le nombre de comptes résidentiels en date du 31 mai par code de district capital. Un code de district capital (il en existe environ 248 dans la province) est une unité administrative utilisée par la B.C. Hydro, qui correspond à une municipalité là où des municipalités existent. Aux termes d'une entente, les deux principaux fournisseurs d'électricité dans la province paient à chaque municipalité un certain pourcentage des recettes annuelles perçues auprès des clients dans cette municipalité, ce qui tient lieu d'impôt foncier. Par conséquent, ces sociétés, comme la B.C. Hydro, établissent leur système de comptabilité en fonction des clients résidant dans les municipalités visées. En outre, la B.C. Hydro cherche à maintenir une correspondance étroite entre les limites des districts capitaux et celles des districts scolaires.

La deuxième forme de présentation indique pour chacun des comptes résidentiels, qui sont au nombre d'un million et plus, le code postal de l'adresse de facturation. Ce deuxième fichier de données permet de convertir facilement les comptes de l'Hydro en unités géographiques autres que les municipalités et les districts scolaires, au moyen du code postal.

4. AVANTAGES DES DONNÉES DE L'HYDRO DANS UN MODÈLE DE RÉGRESSION

Des tests empiriques portant sur les deux sources de données différentes, les comptes de l'Hydro (H) et les comptes d'allocations familiales (F), ont été effectués par la production d'estimations de la population pour 1981, avec chacune prise séparément et avec les deux ensemble. Les coefficients de régression utilisés étaient fondés sur les périodes 1971/1976 et 1976/1981, et l'année de base était 1976. Les résultats ont été comparés aux données du recensement de 1981, et les erreurs absolues moyennes en pourcentage (EAMP) ont été calculées (voir tableaux 1 et 2).

Comme on peut voir au tableau 1, les estimations de la population fondées sur les données de l'Hydro produisent, en moyenne, des pourcentages d'erreur inférieurs à ceux des estimations fondées sur les comptes d'allocations familiales. Si on examine de plus près le tableau 1, on constate que l'amélioration de la précision des estimations touche en presque totalité les estimations relatives aux régions d'une population inférieure à 4,000 habitants. Cette observation est confirmée par le tableau 2 où l'on peut voir que, en termes statistiques, il existe une différence importante sur le plan de la précision entre les estimations fondées sur les comptes de l'Hydro et celles fondées sur les comptes d'allocations familiales pour les régions de petite taille (c.-à-d. moins de 4,000 habitants).

On peut évaluer l'effet marginal de l'adjonction d'un autre indicateur de population à la méthode de corrélation des différences en examinant la variation de la précision de l'estimation selon qu'on utilise ou non cet indicateur. D'après les tableaux 1 et 2, il semblerait que l'adjonction des données de l'Hydro améliore du point de vue statistique la précision de

Utilisation des comptes de l'Hydro dans le modèle d'estimation par régression en Colombie-Britannique¹

DONALD G. McRAE²

RÉSUMÉ

La précision des estimations régionales de la population qui sont établies à l'aide d'un modèle de régression repose foncièrement sur la capacité des indicateurs sélectionnés de traduire avec exactitude les changements démographiques. Par conséquent, il est important de connaître préalablement les caractéristiques des données administratives pouvant servir d'indicateurs de la population dans un modèle de régression. Ce document présente sommairement les avantages et les inconvénients de l'utilisation des comptes résidentiels de l'Hydro dans le modèle d'estimation par régression de la population de la Colombie-Britannique.

MOTS CLÉS: Estimations régionales de la population; méthode de régression; méthode de corrélation des différences; indicateurs de population; comptes de l'Hydro; prestataires d'allocation familiales.

1. INTRODUCTION

Le Bureau central de la statistique produit des estimations postcensitaires de la population pour diverses unités géographiques de la Colombie-Britannique (par ex. les municipalités, les districts sanitaires locaux, les divisions de recensement, les régions de la GRC, et autres). Les estimations courantes de la population sont produites pour ces zones infraprovinciales à l'aide d'une méthode de régression, la méthode de corrélation des différences (MCD). Une description détaillée de cette méthode a été présentée dans des documents antérieurs (Bureau central de la statistique, 1982, McRae 1985). Les données utilisées comme indicateurs de la population sont le nombre de prestataires d'allocations familiales (F), et(ou) le nombre de comptes résidentiels de l'Hydro (H). Les caractéristiques de cette deuxième source de données, les comptes résidentiels de l'Hydro, tels qu'ils sont utilisés dans le modèle de la C.-B., sont examinées dans le reste de cet exposé.

2. SOURCES DE DONNÉES

Les données des comptes résidentiels de l'Hydro à l'intérieur de la province proviennent de neuf administrations et sociétés différentes :

Organisation	% de l'ensemble des comptes de l'Hydro (1985)
(1) British Columbia Hydro	90.8
(2) West Kootenay Power and Light	4.7
(3) Princeton Light and Power Co.	0.2
(4) Ville de Kelowna	0.8
(5) Ville de Penticton	0.8
(6) District de Summerland	0.3
(7) Ville de New Westminster	1.7
(8) Ville de Grand Forks	0.1
(9) Ville de Nelson	0.6

¹ Exposé présenté à la rencontre du Comité fédéral-provincial de la démographie, les 28 et 29 novembre 1985. D.G. McRae, Bureau central de la statistique, ministère du développement de l'industrie et de la petite entreprise de la Colombie-Britannique, 2^e étage, 1405, Douglas Street, Victoria, C.-B., Canada, V8W 3C1.
Les opinions exprimées dans ce document sont celles de l'auteur et ne représentent pas nécessairement le point de vue du gouvernement de la Colombie-Britannique.

4. CONCLUSION

Notre expérience de l'exploitation des dossiers de l'assurance-maladie s'est révélée très positive. Les principaux avantages sont l'utilisation des chiffres dans un modèle des postes appliqué à la production d'estimations intraprovinciales de la population ainsi que l'obtention d'excellents ratios de distribution selon l'âge et le sexe et des tendances uniformes. Les coûts liés au développement du système d'enregistrement des données démographiques ne sont pas jugés excessifs compte tenu de ces avantages. Le Bureau est à la disposition de tout organisme provincial qui envisage de tirer ainsi profit des fichiers de l'assurance-maladie et qui aimerait traiter de la question plus en détail et obtenir des renseignements supplémentaires comme les clichés d'enregistrement et les coûts de traitement des systèmes.

Tableau 4

Comparaisons des chiffres des recensements municipaux de l'Alberta
et des estimations du B.S.A., certaines municipalités

Municipalité		1982			1983			1984		
	Estima- tions Recen- sment du Bureau ^a cipal	Ecart %	Esti- tions Recen- sment du Bureau ^a cipal	Ecart %	Esti- tions Recen- sment du Bureau ^a cipal	Ecart %	Esti- tions Recen- sment du Bureau ^a cipal	Ecart %	Recen- sment du Bureau ^a cipal	
Airdrie	9,450	-5.3	9,830	10,430	-5.8	10,080	--	--	--	
Brooks	9,640	--	9,790	--	--	9,510	--	--	--	
Calgary	614,930	-1.3	622,510	620,690	0.3	615,140	-0.8	-0.8	-0.8	
Camrose	12,880	0.6	12,970	--	--	13,070	2.5	2.5	2.5	
Crowsnest Pass	7,490	-1.1	7,530	--	--	7,350	--	--	--	
Drayton Valley	5,120	5.2	5,200	--	--	5,310	7.9	7.9	7.9	
Drumheller	6,660	--	6,700	6,670	0.4	6,620	--	--	--	
Edmonton ^b	550,930	-0.1	557,400	560,090	-0.5	551,140	--	--	--	
Edson ^b	6,110	-2.9	6,220	--	--	6,080	-14.5	-14.5	-14.5	
Fort McMurray	32,930	-1.9	33,600	34,490	-2.6	35,150	-0.6	-0.6	-0.6	
Fort Saskatchewan	12,530	0.6	12,650	12,470	1.4	12,620	--	--	--	
Grande Prairie	24,650	--	24,910	24,080	3.5	25,370	3.9	3.9	3.9	
Hinton	8,820	0.0	8,980	8,830	1.8	8,950	0.6	0.6	0.6	
Innisfail	5,420	-0.4	5,460	--	--	5,440	0.0	0.0	0.0	
Lacombe	5,810	1.5	5,850	5,850	5,950	-1.8	5,850	--	--	
Leduc	12,880	--	13,010	--	--	13,290	--	--	--	
Lethbridge ^b	55,440	-1.9	55,900	58,090	-3.8	57,500	--	--	--	
Medicine Hat ^b	41,070	--	41,440	42,270	0.7	41,540	--	--	--	
Peace River	6,080	--	6,150	--	--	6,250	--	--	--	
Ponoka	5,310	--	5,310	--	--	5,280	--	--	--	
Red Deer	48,450	-0.2	49,230	50,260	-2.0	50,860	-0.4	-0.4	-0.4	
Spruce Grove	11,080	2.7	11,410	11,310	0.9	11,550	-0.1	-0.1	-0.1	
St. Albert	33,170	0.6	33,740	35,030	-3.7	34,840	-1.9	-1.9	-1.9	
Stettler	5,180	--	5,220	--	--	5,300	--	--	--	
Taber	6,140	--	6,210	--	--	6,360	-0.4	-0.4	-0.4	
Vegreville	5,280	0.6	5,290	--	--	5,390	--	--	--	
Wetaskiwin	9,880	-0.2	9,990	10,020	-0.3	10,080	--	--	--	
Whitecourt	5,710	--	5,840	--	--	5,710	--	--	--	

^a Données expérimentales.

^b Une fusion a eu lieu entre 1982 et 1984.

Note: "--" indique qu'il n'y a pas eu de recensement municipal.

Source: Municipal Affairs d'Alberta, recensements municipaux 1982-1984; estimations du Bureau de la statistique de l'Alberta.

Tableau 3
Comparaisons des chiffres du recensement du Canada et des estimations de la population du B.S.A., certaines municipalités de l'Alberta

Municipalité	Recen- ment de 1976	Estimations du bureau ^a			Écart absolu moyen
		Accrois- sement naturel ^b 1976-1981	Migration nette 1976-1981	Popu- lation 1981	Recense- ment de 1981

Airdrie	1,410	580	5,090	7,070	8,410	-15.9	3.2
Brooks	6,340	730	2,370	9,440	9,420	0.2	0.0
Calgary	469,920	30,310	93,760	593,990	592,740	0.2	0.0
Camrose	10,100	150	2,570	12,830	12,570	2.1	0.4
Crowsnest Pass	5,250	40	-410	4,880	7,310	-33.2	6.6
Drayton Valley	4,300	530	1,760	6,590	5,040	30.8	6.2
Drumheller	6,150	20	220	6,390	6,510	-1.8	0.4
Edmonton	461,360	27,900	51,240	540,510	532,250	1.6	0.3
Edson	4,040	510	2,490	7,040	5,840	20.5	4.1
Fort McMurray	15,420	2,900	14,140	32,460	31,000	4.7	0.9
Fort Saskatchewan	8,300	800	2,660	11,760	12,170	-3.4	0.7
Grande Prairie	17,630	1,970	6,300	25,900	24,260	6.8	1.4
Hinton	6,730	760	-820	6,670	8,340	-20.0	4.0
Innisfail	2,900	230	1,930	5,060	5,250	-3.6	0.7
Lacombe	3,890	150	1,210	5,240	5,590	-6.3	1.3
Leduc	8,580	920	3,430	12,930	12,470	3.7	0.7
Leithbridge	46,750	2,070	4,400	53,220	54,070	-1.6	0.3
Medicine Hat	32,810	1,770	6,010	40,590	40,380	0.5	0.1
Peace River	4,840	580	970	6,390	5,910	8.1	1.6
Ponoka	4,640	-10	530	5,160	5,220	-1.1	0.2
Red Deer	32,180	2,300	11,790	46,270	46,390	-0.3	0.1
Spruce Grove	6,910	1,110	4,710	12,730	10,330	23.2	4.6
St. Albert	24,130	2,360	6,670	33,160	32,000	3.6	0.7
Stettler	4,180	500	580	5,270	5,140	2.5	0.5
Taber	5,300	320	410	6,020	5,990	0.5	0.1
Vegreville	4,160	80	860	5,090	5,250	-3.0	0.6
Wetaskiwin	6,750	300	2,440	9,490	9,600	-1.1	0.2
Whitehorse	3,880	600	1,150	5,630	5,590	0.7	0.1
Alberta	1,838,040	123,020	274,580	2,235,630	2,237,720	-0.1	0.0

^a Données expérimentales.

^b Accroissement naturel désigne le nombre des naissances moins le nombre de décès.

Note: Les chiffres ayant été arrondis, leur somme peut ne pas correspondre aux totaux.

Source: Statistique Canada, recensements de 1976 et 1981; estimations du Bureau de la statistique de l'Alberta.

Tableau 2
Comparaisons des chiffres du recensement du Canada et des estimations de la population du B.S.A., divisions de recensement de l'Alberta

Estimations du bureau^a

Division de recensement	Recensement de 1976	Accroissement naturel ^b de 1976-1981	Migration nette 1976-1981	Croissance 1976-1981	Population 1981
-------------------------	---------------------	---	---------------------------	----------------------	-----------------

1	47,000	2,730	6,080	8,810	55,810
2	96,980	6,120	7,190	13,310	110,290
3	32,870	2,310	100	2,410	35,280
4	12,140	490	-520	-30	12,110
5	35,460	1,820	790	2,610	38,070
6	524,570	33,860	107,540	141,400	665,970
7	37,820	2,010	-10	2,000	39,820
8	95,400	6,140	20,860	27,000	122,400
9	19,850	1,040	200	1,240	21,090
10	67,230	1,650	8,550	10,200	77,430
11	632,830	43,880	90,880	134,760	767,590
12	63,130	6,470	16,130	22,600	85,730
13	46,300	2,040	4,320	6,360	52,660
14	19,450	2,200	2,430	4,630	24,080
15	107,010	10,260	10,040	20,300	127,310
Alberta	1,838,040	123,020	274,580	397,600	2,235,640

Division de recensement	Recensement 1981	Ecart		Ecart absolu moyen
		Nombre	%	

1	55,360	450	0.81	0.16
2	110,470	-180	-0.16	0.03
3	35,640	-360	-1.01	0.20
4	12,120	-10	-0.08	0.02
5	38,430	-360	-0.94	0.19
6	668,680	-2,710	-0.41	0.08
7	40,030	-210	-0.52	0.10
8	123,690	-1,290	-1.04	0.21
9	21,630	-540	-2.50	0.50
10	78,390	-960	-1.22	0.24
11	762,080	5,510	0.72	0.14
12	84,220	1,510	1.79	0.36
13	53,690	-1,030	-1.92	0.38
14	24,650	-570	-2.31	0.46
15	128,640	-1,330	-1.03	0.21
Alberta	2,237,720	-2,080	-0.09	0.02

^a Données expérimentales.

^b Accroissement naturel désigne le nombre des naissances moins le nombre de décès.

Note: Les chiffres ayant été arrondis, leur somme peut ne pas correspondre aux totaux.

Source: Statistique Canada, recensements de 1976 et 1981; estimations du Bureau de la statistique de l'Alberta.

Le tableau 2 présente les résultats d'une comparaison, au niveau des divisions de recensement, entre les chiffres du recensement de 1981 et les estimations relatives à la même année produites, à l'aide de la méthode, à partir des chiffres de la population dénombrée au recensement de 1976. Pour treize des quinze divisions, les estimations se situaient à $\pm 2.0\%$ des chiffres du recensement de 1981. Seules les deux plus petites DR (9 et 14) affichaient pour cette période de cinq ans un écart plus grand que 2.0% . Les écarts absolus moyens (ou écarts annuels moyens) ne dépassaient pas 0.5% dans l'ensemble des divisions du recensement.

Les estimations de la population de vingt-huit municipalités ont été comparées aux chiffres du recensement de 1981 ainsi qu'aux données tirées des recensements municipaux menés entre 1982 et 1984 (voir tableaux 3 et 4). Les comparaisons avec les chiffres du recensement fédéral ont révélé que les estimations de dix-neuf des vingt-huit municipalités présentaient un écart absolu moyen de $\pm 1.0\%$. Seulement six municipalités affichaient un écart annuel supérieur à 2.0% . Les comparaisons avec les recensements municipaux menés entre 1982 et 1984 ont donné vingt-deux cas d'écarts en deçà de $\pm 1.0\%$, quatorze cas d'écarts variant entre $\pm 1.0\%$ et $\pm 3.0\%$ et des écarts supérieurs à $\pm 3.0\%$ dans neuf cas.

Les résultats de l'estimation sont dans l'ensemble satisfaisants et encourageants. La production des chiffres à partir des dossiers d'inscription de l'AHCIIP ainsi que la méthode des composantes utilisées ont amélioré la précision des estimations de la population et ont ouvert de nouvelles avenues pour la production d'estimations sur les petites régions géographiques définies par les utilisateurs. Le Bureau continuera de rechercher des façons d'améliorer les totalisations fondées sur les dossiers de l'AHCIIP (certaines concernant l'adoption de nouvelles procédures administratives pour l'AHCIIP). La méthode d'estimation de la population fera elle-même l'objet de perfectionnements à mesure que le Bureau aura accès à de nouvelles données ou nouvelles techniques.

3.3 Résumé des avantages et désavantages de l'utilisation des dossiers de l'AHCIIP

L'utilisation des chiffres sur les bénéficiaires de l'assurance-maladie pour la production d'estimations régionales de la population comporte un certain nombre d'avantages et de désavantages.

Avantages

- a) Les données sur les enrégistrement de l'AHCIIP s'appliquent à l'ensemble des habitants de l'Alberta.
- b) Les délais d'inscription semblent aléatoires et n'influent pas sur les distributions ou les tendances observées.
- c) Les données sont diffusées rapidement et fréquemment.
- d) Le fichier contient des renseignements socio-économiques sur les bénéficiaires (notamment l'âge, le sexe et l'état civil), ce qui permet la production d'estimations détaillées sur la population.

Désavantages

- a) Le lieu de résidence déterminé en fonction du code postal peut entraîner certaines inexactitudes.
- b) Les personnes inscrites aux dossiers de l'AHCIIP peuvent quitter le système, notamment à la suite d'un décès ou d'un départ vers une autre province, sans que les bureaux de l'AHMC en aient été avisés, ce qui entraîne un surdénombrement.
- c) Les procédures administratives peuvent être la cause d'écarts/d'inexactitudes dans les comptes des bénéficiaires de l'assurance-maladie de l'Alberta.

actuellement en se basant sur les mouvements interprovinciaux des bénéficiaires d'allocation familiales. (Comme Statistique Canada utilise également les fichiers des allocations familiales, les estimations du Bureau sont généralement très proches des estimations produites par l'agence fédérale.)

Pour améliorer les estimations intraprovinciales de la migration et assurer leur compatibilité avec les estimations provinciales, on a introduit un facteur de correction dans les estimations. Il s'agit d'établir le rapport entre les chiffres de migration nette des bénéficiaires de l'assurance-maladie pour une région donnée et les chiffres de migration nette des bénéficiaires de la province calculée à partir des estimations trimestrielles de la population produites par le Bureau. L'équation mathématique s'écrit:

$$AMIG_t = \frac{HMIG_t}{HMIG_a} \times PMIG$$

où:

$AMIG_t$ = migration nette corrigée dans la région t

$HMIG_t$ = chiffres de migration nette des bénéficiaires de l'assurance-maladie pour la région t

$HMIG_a$ = chiffres de migration nette des bénéficiaires de l'assurance-maladie pour l'Alberta

$PMIG$ = migration provinciale nette estimée à partir des estimations trimestrielles de la population.

Ces estimations corrigées de la migration ($AMIG$) sont ensuite utilisées pour la production d'estimations de la population.

b) Estimations régionales de la population des petites régions

L'estimation corrigée de la migration nette ($AMIG$) calculée pour chaque région est introduite dans une équation comprenant les composantes de la croissance démographique (naissances, décès et migration):

$$P_t = P_{t-1} + (B_t - D_t) + AMIG_t$$

Où:

P_t = population estimée dans la région t au moment t

P_{t-1} = population dans la région t au moment $t-1$

3.2 Évaluation des estimations pour les petites régions

En utilisant la méthode décrite ci-dessus, le Bureau a produit des estimations de la population pour les quinze divisions de recensement de l'Alberta et vingt-huit municipalités comptant plus de 5,000 habitants. Les résultats sont jusqu'ici très encourageants.

3. UTILISATION DES CHIFFRES SELON L'AHCIPI POUR LA PRODUCTION D'ESTIMATIONS DE LA POPULATION DES PETITES RÉGIONS

Depuis près de dix ans, le Bureau produit des estimations intercensitaires de la population de l'Alberta et de ses divisions de recensement. Au cours de cette période, diverses méthodes ainsi que diverses sources de données ont été analysées et quelques-unes utilisées pour améliorer la qualité des estimations. À ce jour, on a obtenu des résultats très satisfaisants avec la méthode des composantes, en utilisant comme données d'entrée les chiffres sur les bénéficiaires de l'assurance-maladie. Ces données ont servi à calculer la structure par âge et par sexe de la population de l'Alberta, au niveau de la province et des divisions de recensement, et à produire des estimations de la population à ces mêmes niveaux. Elles ont également été utilisées récemment dans des tests visant à déterminer leur utilité dans la production d'estimations de la population au niveau des subdivisions de recensement.

3.1 Méthode d'estimation

La méthode d'estimation utilisée par le Bureau pour produire des estimations infraprovinciales de la population comprend deux volets: l'estimation de la population migrante et la production des chiffres de population.

a) Estimation de la population migrante à partir des chiffres sur les bénéficiaires de l'assurance-maladie

Le Bureau tire ses données de trois fichiers administratifs: les chiffres fondés sur les dossiers de l'AHCIPI; les données sur les naissances fournies par la statistique de l'état civil de l'Alberta; et les données sur les décès également fournies par la statistique de l'état civil de l'Alberta. Les données des trois sources servent au calcul de la migration nette. Fondamentalement, pour toute petite région donnée, on mesure l'accroissement du nombre des bénéficiaires de l'assurance-maladie en faisant la différence des totaux enregistrés entre le moment t et le moment $t-1$. De ce résidu, on retranche l'accroissement naturel de la population de la région (les naissances moins les décès) pour calculer les entrées (ou les sorties) nettes d'individus, c'est-à-dire la migration nette. L'équation mathématique prend la forme:

$$HMIG = [(HC_t - HC_{t-1}) - (B - D)]$$

où:

- $HMIG$ = estimation de la migration nette des bénéficiaires de l'assurance-maladie entre le moment t et le moment $t-1$
- HC_t = nombre de bénéficiaires de l'assurance-maladie au moment t
- HC_{t-1} = nombre des bénéficiaires de l'assurance-maladie au moment $t-1$
- B = naissances totales au cours de l'intervalle t à $t-1$
- D = décès totaux au cours de l'intervalle t à $t-1$.

Toutefois, les problèmes de surdénombrement et de sous-dénombrement énoncés dans la partie 2 s'appliquent également à cette méthode d'estimation de la population migrante des bénéficiaires de l'assurance-maladie. Aussi, bien que des estimations infraprovinciales puissent être obtenues avec l'application d'une telle méthode, une fois ramenées à un niveau d'agrégation provincial, elles sont moins fiables que les estimations que le Bureau produit

À mesure que la taille de la région géographique diminue, les chiffres selon l'AHCIIP deviennent moins fiables; les distributions selon l'âge et le sexe, bien que moins précises, demeurent adéquates et l'uniformité des tendances (évolution des chiffres dans le temps) continue à présenter un degré élevé de corrélation, à quelques exceptions près. Les limites que présentent les chiffres basés sur les dossiers de l'AHCIIP en tant qu'indicateurs de la population peuvent essentiellement être attribuées à l'une des deux grandes causes suivantes: a) les inexactitudes imputables aux procédures administratives de l'AHCIIP; b) l'utilisation des codes postaux.

a) *Procédures administratives de l'AHCIIP*

- 1) Comme il s'agit d'un régime d'assurance, la préoccupation première est d'offrir le service à l'ensemble de la population. En conséquence, pour assurer l'universalité du régime, on consacre davantage d'efforts à l'entrée des gens dans le système qu'au contrôle des sorties. C'est pourquoi le nombre des gens inscrits dépasse la population réelle de la province.
- 2) On utilise les adresses postales plutôt que les adresses résidentielles, ce qui complique l'attribution des codes géographiques. On note des écarts importants dans les régions comprenant une population rurale importante qui vit à proximité d'un centre urbain et ramasse son courrier dans le centre urbain. Le cas échéant, il y a surdénombrement dans la plupart des régions urbaines et sous-dénombrement dans les régions rurales.
- 3) Des données incomplètes et inexacts, particulièrement celles touchant aux codes postaux, rendent difficile la production de statistiques régionales à cause du sous-dénombrement.
- 4) Les délais de déclaration et d'enregistrement des données influent sur les totalisations. Dans la plupart des cas, il faut compter de trois à six mois avant qu'un individu n'entre dans le système (naissance, immigration), mais il faut encore plus de temps pour retirer un dossier actif du système (décès, émigration). Toutefois, ces délais sont difficiles à calculer et varient considérablement selon les circonstances.

b) *Codes postaux*

- 1) Le code postal délimite la zone du service de livraison (l'endroit où une personne ramasse son courrier) qui n'est pas nécessairement le lieu de résidence. Ce facteur limite la précision de la répartition géographique des enregistrements de l'AHCIIP. Comme on l'a vu, cela entraîne un problème de démarcation entre les régions urbaines et rurales.
- 2) Le code postal à six chiffres ne suffit pas toujours pour déterminer la zone du service de livraison. Pour une adresse plus exacte, il est parfois nécessaire de préciser également la route rurale, le service de banlieue ou la case postale.
- 3) Les codes postaux n'ont pu donner des niveaux d'aggrégation adéquats, surtout dans les régions rurales. À titre d'exemple, il y a environ 363 subdivisions de recensement en Alberta, mais le FCCP du Bureau peut en extraire seulement 324.

Les problèmes énoncés ci-dessus ont empêché la diffusion des chiffres basés sur les dossiers de l'AHCIIP en tant qu'approximations de la population réelle. Il est vrai qu'on a obtenu un degré élevé de précision dans certaines régions, mais dans d'autres, les chiffres fondés sur les dossiers de l'AHCIIP étaient peu représentatifs. Cependant, compte tenu des relations étroites entre les distributions selon l'âge et le sexe des bénéficiaires de l'assurance-maladie et les distributions fondées sur les chiffres du recensement du Canada et compte tenu également de l'uniformité des tendances dans le temps, les chiffres basés sur les dossiers de l'AHCIIP ont été inclus dans la méthode qu'utilise le Bureau pour produire des estimations de la population (méthode expliquée dans la partie qui suit).

Tableau 1
Comparaisons des chiffres selon l'AHCIIP et des chiffres du recensement du Canada, divisions de recensement de l'Alberta

Année											
Division de recensement	Chiffres du recensement	1976					1981				
		Chiffres selon l'AHCIIP	Ecart en pourcentage	Ecart en chiffres réels	Chiffres du recensement	Chiffres selon l'AHCIIP	Ecart en pourcentage	Ecart en chiffres réels			
1	46,990	45,789	-2.56	-1,201	55,375	55,748	0.67	373			
2	96,995	97,229	0.24	234	110,477	111,567	0.99	1,090			
3	32,898	33,884	3.00	986	35,652	36,463	2.27	811			
4	12,130	12,101	-0.24	-29	12,119	12,038	-0.67	-81			
5	35,424	35,656	0.65	232	38,382	38,457	0.20	75			
6	524,554	538,432	2.65	13,878	668,682	699,999	4.68	31,317			
7	37,866	38,235	0.97	369	40,071	40,359	0.72	288			
8	95,384	95,063	-0.34	-321	123,642	124,666	0.83	1,024			
9	19,903	21,832	9.69	1,929	21,670	23,338	7.70	1,668			
10	67,171	67,168	0.00	-3	78,417	78,532	0.15	115			
11	632,909	646,799	2.19	13,890	762,041	796,884	4.57	34,843			
12	63,129	62,011	-1.77	-1,118	84,221	86,183	2.33	1,962			
13	46,305	47,258	2.06	953	53,701	54,282	1.08	581			
14	19,386	21,039	8.53	1,653	24,635	25,991	5.50	1,356			
15	106,993	111,678	4.38	4,685	128,639	134,451	4.52	5,812			
Non connue ^a		48,462			19,279						
Alberta		1,838,037	1,922,636	4.60	84,599	2,237,724	4.49	100,513			

^a Cette rubrique englobe les totalisations obtenues pour tous les cas pour lesquels il n'y avait pas d'identification d'adresse.
Source: Recensements de Statistique Canada de 1976 et 1981; Alberta Health Care Insurance Plan

c) Au niveau des subdivisions de recensement unifiées (SRU), l'écart entre les données sur les bénéficiaires de l'assurance-maladie et les chiffres du recensement $\pm 10\%$ dans cinquante des soixante et onze SRU. L'écart le plus important est de -56.5% (district municipal 135).

Pour la plupart des régions problèmes, on a constaté que le comté, district municipal ou district d'amélioration locale était situé à proximité d'un grand centre urbain. Aucune anomalie précise n'a été relevée lors de la vérification des distributions selon l'âge et le sexe, quoique les relations n'étaient pas aussi étroites que dans le cas des comparaisons au niveau de la province et des divisions de recensement.

d) Au niveau des subdivisions de recensement (SDR), les chiffres préliminaires révélaient des écarts variant entre -100% et $+95.5\%$ entre les chiffres selon l'AHCIIP et les chiffres du recensement de 1981. On a donc décidé d'utiliser à ce niveau d'agrégation les vingt-neuf régions les plus importantes comptant plus de 5,000 habitants. Les six SDR les plus grandes (Edmonton, Calgary, Lethbridge, Medicine Hat, Red Deer et St. Albert) affichaient un surdénombrement variant de 3% à 9% . Pour huit autres SDR, l'écart se situait à $\pm 20\%$ et dans seize autres cas, la variation était légèrement supérieure à $\pm 20\%$. Une fois de plus, aucune anomalie précise n'a été décelée dans la distribution selon l'âge et le sexe, quoique les écarts étaient plus importants qu'à des niveaux d'agrégation plus hauts. À preuve, vingt-sept des vingt-neuf SDR observées affichaient un niveau élevé d'uniformité des tendances.

sur bande constitue une liste partielle puisqu'on y a biffé tous les noms et autres identifiants de façon à préserver le caractère confidentiel des renseignements sur les individus. Les renseignements types contenus dans le fichier sont l'adresse, le code postal, les dates d'inscription et d'annulation, l'âge et le sexe. (On peut obtenir sur demande une description détaillée du cliché d'enregistrement).

L'unité de déclaration du fichier de l'AHICIP est l'inscription. Chaque inscription peut comprendre jusqu'à vingt-cinq individus: un requérant (habituellement la personne qui paie les primes) et jusqu'à concurrence de vingt-quatre personnes à charge. Il y a actuellement environ 1,7 million de dossiers d'inscription actifs qui représentent environ 2,6 millions d'individus. De plus, le fichier est chronologique et englobe toutes les personnes protégées en vertu de l'AHICIP depuis son entrée en vigueur en 1969.

Le traitement du fichier se fait en quatre étapes:

- a) Vérification – notation ou correction des erreurs en fonction de critères de contrôle précis.
- b) Épuration – sélection des individus actifs à partir du fichier des données corrigées à l'étape d).

- c) Codification – appariement des codes postaux du fichier épuré et du fichier de conversion des codes postaux du Bureau (FCCP) et attribution de codes de référence géographique aux enregistrements de l'AHICIP.

- d) Agrégation – totalisation des individus des deux sexes par année d'âge et pour chaque code postal, à partir du fichier codifié. Cette opération permet de réduire le nombre des enregistrements/individus d'environ 2,6 millions à moins de 120,000 et donc de diminuer sensiblement les coûts de traitement informatique subséquents.

Le fichier agrégatif sert à la production de chiffres de la population selon l'âge et le sexe pour toute région géographique pouvant être définie à l'aide des 60,000 codes du FCCP de l'Alberta.

2.2 Évaluation des résultats

Pour évaluer les chiffres sur les bénéficiaires de l'assurance-maladie, on les a comparés aux chiffres du recensement de la population du Canada de 1976 et de 1981. Comme les données du fichier de 1981 ont été jugées plus précises que celles du fichier de 1976, on s'est davantage fondé sur les comparaisons avec les chiffres du recensement de 1981. Les données des recensements municipaux ont été utilisées comme deuxième base de comparaison, bien qu'elles ne soient pas jugées aussi fiables dans l'ensemble que les chiffres du recensement du Canada. Toutefois, les données des recensements municipaux ont permis de mieux saisir l'importance des variations ainsi que les distributions relatives selon l'âge et le sexe et l'évolution des tendances (croissance ou déclin). On a en outre établi des comparaisons avec les estimations intercenitaires de la population produites par le Bureau et par Statistique Canada.

Principales constatations

- a) À l'échelle provinciale, les chiffres basés sur les dossiers de l'AHICIP surestiment d'environ 3,5% à 4,5% aussi bien les chiffres du recensement du Canada que les chiffres des recensements municipaux. Les distributions selon l'âge et le sexe sont plus précises et les coefficients de corrélation témoignent de l'uniformité des tendances (surestimation/sous-estimation) dans le temps.
- b) Au niveau des divisions de recensement (DR), on enregistre des écarts allant de -2,6% à 9,7% entre les chiffres de l'AHICIP et ceux du recensement du Canada (voir tableau 1). La variation est sensiblement la même dans le cas des comparaisons avec les estimations intercenitaires de la population. Comme pour les données provinciales, les distributions selon l'âge et le sexe et l'uniformité des tendances se sont révélées d'une fiabilité élevée.

Utilisation des dossiers de l'assurance-maladie de l'Alberta pour la production d'estimations régionales de la population¹

F. AHMAD, R. CHOW, O. DEVRIES,
A. HASHMI, et M. MARCOGLIESE²

RÉSUMÉ

Ce document traite de l'utilisation des fichiers administratifs du régime d'assurance-maladie de l'Alberta et de la statistique de l'état civil pour la production d'estimations de la population. Les données obtenues et comparées avec les données du recensement montrent que les fichiers d'assurance-maladie peuvent servir à la production d'estimations précises à l'échelle de la province et au niveau de petites régions telles que les divisions de recensement et les municipalités.

MOTS CLÉS: Fichiers administratifs; méthode des composantes; petites régions; migration nette résiduelle.

1. HISTORIQUE

Du milieu à la fin des années 70, la province de l'Alberta a connu une croissance économique rapide, en raison surtout de l'activité du secteur du pétrole et du gaz, qui a entraîné un accroissement marqué de la population. Les divers ordres de gouvernement ont exigé des données à jour sur l'importance et la répartition régionale de cette croissance démographique, afin de pouvoir offrir à la population des biens et services adéquats. Face à ce besoin, on a jugé que les recensements menés tous les cinq ans par le gouvernement fédéral n'étaient ni assez fréquents, ni suffisamment à jour puisque les chiffres du recensement sont publiés de douze à dix-huit mois après l'année de référence. En conséquence, des organismes provinciaux, et tout particulièrement le Bureau de la statistique de l'Alberta (désigné ci-après le Bureau), ont entrepris des recherches en vue de trouver de nouvelles sources de données, disponibles rapidement, sur la population.

Après avoir examiné un certain nombre de possibilités, le Bureau a procédé à l'évaluation des données administratives sur l'assurance-maladie tirées des fichiers de l'Alberta Health Care Insurance Plan (AHCHIP) en vue de la production de statistiques démographiques. Le présent document souligne les travaux entrepris par le Bureau pour faire en sorte que les données tirées des dossiers de l'AHCHIP puissent servir à des estimations de population pour des petites régions.

2. PRODUCTION DE CHIFFRES SUR LES BÉNÉFICIAIRES DE L'ASSURANCE-MALADIE À PARTIR DES DOSSIERS DE L'AHCHIP

La présente partie décrit brièvement la nature des dossiers de l'AHCHIP et évalue les utilisations produites à partir de l'information qui y est contenue.

2.1 Elaboration des données sur les bénéficiaires de l'assurance-maladie

Chaque trimestre, le Bureau reçoit sur bande informatisée un certain nombre d'enregistrements tirés du système d'inscription et de facturation de l'AHCHIP. Le fichier

¹ Version condensée du document présenté à la rencontre du Comité fédéral-provincial de la démographie tenue les 26 et 27 novembre 1985 à Ottawa.
² F. Ahmad, R. Chow, O. Devries, A. Hashmi et M. Marcogliese, Bureau de la statistique de l'Alberta, Sir Frederick W. Haultain Building, 9811-109th Street, Edmonton, Alberta, Canada, T5K 0C8.

BIBLIOGRAPHIE

- BUREAU DE LA STATISTIQUE DE L'ALBERTA (1985). Utilisation des dossiers de l'assurance-maladie de l'Alberta pour la production d'estimations régionales de la population. Document présenté aux réunions du Comité fédéral-provincial de la démographie, Ottawa.
- HOAG, ELIZABETH (1984). Estimating Annual Migration for California Counties using Driver's Licence Address Change. Document présenté aux réunions de la Population Association of America, Minneapolis, Minnesota.
- McRAE, DONALD G. (1985). Utilisation des comptes de l'Hydro dans le modèle d'estimation par régression de la population de la Colombie-Britannique. Document présenté aux réunions du Comité fédéral-provincial de la démographie, Ottawa.
- NORRIS, D., et STANDISH, L. (1983). *Rapport technique sur la production d'estimations à partir de dossiers fiscaux*. Division de l'exploitation des données administratives, Statistique Canada.
- SHRYOCK, HENRY, et SIEGEL, JACOB S. (1971). *The Methods and Materials of Demography*. Washington: U.S. Bureau of Census.

Tableau 5

Indice de dissemblance	
Indice de dissemblance	
Année	Migration d'entrée
1979-80	5.61
1980-81	5.54
1981-82	5.26
1982-83	4.41
	4.87
	3.50
	3.82
	4.82
	Migration de sortie

Les valeurs peu élevées de l'indice indiquent que les deux fichiers (des permis de conduire et de l'impôt sur le revenu) produisent des estimations assez semblables des migrations intraprovinciales dans le cas des divisions de recensement en Ontario. Au cours des quatre dernières années, toutefois, le degré de dissemblance augmente pour ce qui a trait à la migration de sortie et diminue pour ce qui a trait à la migration d'entrée.

5. CONCLUSION ET SOMMAIRE

Cet exposé présente une comparaison des estimations des migrations intraprovinciales produites à partir du fichier des permis de conduire et de celles qui ont été produites à partir du fichier de l'impôt sur le revenu. Les deux fichiers donnent des mesures raisonnablement bonnes des migrations intraprovinciales pour les divisions de recensement en Ontario. Tandis que les données du fichier des permis de conduire semblent avoir produit de meilleures estimations de la direction des migrations intraprovinciales, l'utilisation des données du fichier de l'impôt sur le revenu a produit des erreurs en fin de période intercensitaire plus petites pour un nombre légèrement plus grand de comtés et a de plus donné des résultats un peu meilleurs pour certaines grandes régions (par exemple, la répartition des migrants dans la conurbation de Toronto/Hamilton). Étant donné leurs avantages respectifs, la meilleure méthode semble être de combiner l'utilisation des deux sources de données. Il faut également noter que l'évaluation a porté sur trois années seulement, c'est-à-dire de 1979 à 1981. Il est impossible d'évaluer de façon plus précise la qualité des estimations produites à partir de ces deux fichiers tant que les données du recensement de 1986 ne seront pas connues.

Pour améliorer davantage la qualité et les applications des estimations produites à partir des données sur les permis de conduire, nous proposons d'orienter les recherches dans les deux domaines suivants:

- Vérification du facteur de rajustement FA en tenant compte des chiffres réels des migrations dont l'origine et/ou la destination sont inconnues au moyen du codage manuel des adresses.
- Extension de l'utilisation du fichier des données sur les permis de conduire comme source s'ajoutant aux données des allocations familiales et aux données sur le revenu, pour l'estimation des migrations interprovinciales.

Le fichier des permis de conduire tend à surestimer le nombre de migrants de la région métropolitaine de Toronto et à sous-estimer celui des migrants des régions environnantes. Il semble que ce soit l'inverse qui se produit avec le fichier de l'impôt sur le revenu. Pour l'ensemble de la conurbation de Toronto, le fichier des permis de conduire produit de meilleures estimations des migrations intraprovinciales que le fichier de l'impôt sur le revenu qui, en revanche, donne une meilleure répartition des migrations intraprovinciales dans les comtés de cette région.

de conduire sont utilisées, et au nombre de 10, lorsque les données du fichier de l'impôt sur le revenu sont utilisées (tableau 4). Il s'agit d'un cas où les données sur les permis de conduire ont semblé produire des estimations plus fiables que les estimations produites à partir des données du fichier de l'impôt sur le revenu.

4.4 Indice de dissemblance

L'indice de dissemblance a été calculé pour la migration d'entrée et la migration de sortie séparément, étant donné que le sens de la migration nette indiqué par les deux séries d'estimations n'était pas le même dans certains comtés. La valeur de l'indice de dissemblance peut varier de 0 à 100. Cet indice mesure la moitié de la somme des écarts absolus entre les deux répartitions en pourcentage correspondantes et équivaut à la somme des écarts positifs ou à la somme des écarts négatifs (Shryock et Siegel, 19781). La formule générale est:

$$ID = \frac{1}{2} \sum |r_2 - r_1|$$

où r_2 et r_1 sont les pourcentages correspondants dans les deux répartitions.

Tableau 4

Division de recensement où la variation de population est différente de celle indiquée par les données du recensement		
Division de recensement	les données sur l'impôt sur le revenu	les données du recensement
Variation en % d'après		
Bruce	0.11	-1.77
Grey	-0.04	0.17
Hastings	0.12	-1.03
Leeds et Grenville	0.21	-0.32
Niagara	0.15	-0.46
Northumberland	0.31	-0.82
Oxford	0.17	-0.09
Parry Sound	1.20	-0.51
Stormond/Dundas/Glenagarry	0.44	-0.17
Sudbury T.D.	0.41	-2.28
Province	1.60	1.46
Variation en % d'après		
les données sur les permis de conduire	les données du recensement	
Leeds et Grenville	0.02	-0.32
Parry Sound	0.81	-0.51
Thunder Bay	0.42	-0.20
Province	1.60	1.46

4.1 Erreurs en fin de période intercensitaire

Nous avons comparé les deux séries d'estimations de la population des divisions de recensement (l'une produite à partir des données sur les permis de conduire et l'autre, à partir des données sur l'impôt sur le revenu, dans les deux cas pour les migrations intraprovinciales) avec les chiffres de population du recensement de 1981. La différence exprimée en pourcentage entre la population estimative et le chiffre de population du recensement est appelée "erreur en fin de période intercensitaire". Pour 23 des 49 divisions de recensement, l'erreur est plus petite lorsque les données sur les permis de conduire sont utilisées et, pour 26 DR, l'erreur est plus petite lorsque les données sur l'impôt sur le revenu sont utilisées pour l'estimation des migrations intraprovinciales.

4.2 Conurbation de Toronto

La conurbation de Toronto, qui englobe six municipalités régionales (tableau 3), est un cas assez intéressant. Les données sur les permis de conduire produisent une erreur en fin de période intercensitaire plus petite pour la conurbation dans son ensemble et sous-estiment la population des régions situées à l'extérieur de la région métropolitaine de Toronto. Avec les données du fichier de l'impôt sur le revenu, on obtient une erreur en fin de période intercensitaire plus petite pour chacune des divisions de recensement, alors que dans le cas de la conurbation dans son ensemble, l'erreur en fin de période intercensitaire est plus grande que celle qui est calculée à partir des données du fichier des permis de conduire (tableau 3). Pour cette raison, nous avons utilisé les données sur les permis de conduire pour estimer les migrations intraprovinciales des résidents de la conurbation de Toronto dans son ensemble, alors que nous avons réparti les migrants selon chaque division de recensement de la conurbation d'après les données sur le revenu.

4.3 Taux de croissance de la population

Nous avons calculé le taux de variation entre 1979 et 1981 des estimations de la population produites à partir du fichier de l'impôt sur le revenu et celui des estimations produites à partir des données sur les permis de conduire. Les taux de croissance ainsi obtenus ont ensuite été comparés aux taux de croissance indiqués par les données du recensement de 1981. Les taux de croissance obtenus à partir des deux fichiers ne diffèrent pas beaucoup des taux de croissance indiqués par les données du recensement. Les divisions de recensement où la population ne varie pas dans le même sens que le taux de croissance indiqué par les données du recensement sont au nombre de 3, lorsque les données du fichier des permis

Tableau 3

Erreurs en fin de période intercensitaire, conurbation de Toronto

Erreurs en fin de période intercensitaire		
Division de recensement	Fichier impôt sur le revenu	Fichier permis de conduire
M.R. de Durham	-0.21	-0.41
M.R. de Halton	0.45	-0.51
M.R. de Hamilton-Wentworth	0.10	-0.11
M.R. de Peel	-0.63	-1.94
M.R. de Toronto	0.01	1.26
M.R. de York	-0.11	-5.70
Conurbation de Toronto, Total	-0.05	-0.01

3. CONVERSION DU NOMBRE DE CONDUCTEURS
EN NOMBRE DE MIGRANTS

Pour connaître le nombre de migrants, un facteur de rajustement est appliqué au nombre de conducteurs. Ce facteur est calculé de la façon suivante:

$$FA = \frac{\text{Total des mouvements connus et des mouvements inconnus}}{\text{Nombre de mouvements connus}}$$
$$FB = \frac{\text{Population totale}}{\text{Population détenant un permis de conduire}}$$
$$F = FA \times FB.$$

FA tient compte des origines et/ou des destinations non déclarées tandis que FB tient compte des personnes qui ne détiennent pas de permis de conduire. Le facteur de rajustement FA suppose que les comportements migratoires de ceux qui n'ont pas de permis de conduire ne diffèrent pas des comportements migratoires des détenteurs d'un permis de conduire, tout comme le facteur FB suppose que les comportements migratoires des migrants dont les mouvements n'ont pas été précisés ne diffèrent pas des comportements migratoires des migrants dont les mouvements ont été précisés.

4. ESTIMATION DES MIGRATIONS INTRAPROVINCIALES:
FICHIER DES PERMIS DE CONDUIRE ET FICHIER
DE L'IMPÔT SUR LE REVENU

Statistique Canada utilise les changements d'adresse indiqués par les contribuables dans leur déclaration annuelle d'impôt. Le nombre d'enfants est estimé à partir du nombre de personnes à charge déclaré par le contribuable. Comme dans le cas des données sur les permis de conduire, il faut introduire un facteur de rajustement dans les données sur le revenu pour tenir compte des cas de codes postaux non précisés et des cas des personnes qui ne produisent pas de déclaration d'impôt. En outre, certains contribuables inscrivent une adresse postale autre que celle de leur domicile dans leur déclaration d'impôt.

Il a fallu vérifier la qualité des estimations des migrations produites à partir du fichier de l'impôt sur le revenu et celle des estimations des migrations produites à partir du fichier des données sur les permis de conduire. Nous avons utilisé trois mesures pour vérifier l'exactitude relative des estimations des migrations produites à partir de ces deux fichiers; ce sont:

- A. les erreurs en fin de période intercenitaire
- B. la comparaison des taux de croissance
- C. l'indice de dissemblance.

Idealement, les erreurs en fin de période intercenitaire et les taux de croissance devraient être calculés à partir des estimations de la population d'une année de recensement donnée à l'année de recensement suivante. Comme on dispose de données fiables sur les changements d'adresse des détenteurs de permis de conduire seulement depuis 1979 en Ontario, les estimations intercenitaires de la population faites en 1979 et les chiffres estimatifs de la population en 1981 ont servi de base aux calculs. Deux séries d'estimations de la population ont été calculées pour l'estimation des migrations intraprovinciales: une première à l'aide du fichier des adresses des détenteurs de permis de conduire et une deuxième à l'aide du fichier de l'impôt sur le revenu. Toutes les autres composantes, c'est-à-dire les naissances, les décès, les migrations interprovinciales et les migrations internationales sont demeurent les mêmes pour les deux séries d'estimations de la population.

Plus d'un million de changements d'adresse sont enregistrés chaque année. La plupart de ces déménagements se produisent à l'intérieur même des divisions de recensement (c.-à-d. un comté ou une municipalité régionale). Cependant, les migrations nettes inter-comtés s'élèvent en moyenne à 22,000 seulement par année. On constate, d'après le tableau 1, qu'un tiers environ des enregistrés n'indiquent pas le code postal de l'origine et/ou de la destination du migrant.

En Ontario, une personne peut obtenir un permis de conduire dès l'âge de 16 ans. Aussi plus de 75% de la population visée détient un permis de conduire. Par ailleurs, la proportion de personnes âgées et de femmes qui possèdent un permis de conduire est moins élevée que celle des autres catégories de détenteurs (tableau 2).

Tableau 1

Nombre total de migrants et nombre de migrants dont l'origine et/ou la destination n'ont pas été précisées, Ontario, 1975-1985

Année	Nombre de migrations dont l'origine et la destination sont connues (inter-comté et infracomté)	Nombre de migrations dont l'origine et/ou la destination n'ont pas été précisées	Total	Pourcentage de migration non précisées
1979 (année civile)	881,000	0	881,000	0
1979/80	586,000	301,000	887,000	34
1980/81	566,000	306,000	872,000	35
1981/82	617,000	270,000	887,000	30
1982/83	648,000	259,000	907,000	29
1983/84	822,000	320,000	1,142,000	28
1984/85	831,000	330,000	1,161,000	28

Source: Ministère des Transports et des Communications de l'Ontario.

Tableau 2

Pourcentage de la population détenteur un permis de conduire 1981

Âge	Hommes	Femmes	Total
16-19 ans	63.9	36.5	49.1
20-24 ans	92.7	73.0	85.7
25-34 ans	98.6	81.6	90.0
35-44 ans	99.7	79.9	90.0
45-54 ans	96.7	67.8	82.4
55-64 ans	93.2	56.7	74.2
65 ans et plus	73.1	27.4	46.4
Total	90.6	62.2	75.8

Source: Ministère des Transports et des Communications de l'Ontario.

Utilisation de fichiers de données administratives pour la production d'estimations de la migration: une étude de cas du fichier des permis de conduire en Ontario¹

RAGHUBAR D. SHARMA et CHEUK WONG²

RÉSUMÉ

Au Canada, les démographes provinciaux et fédéraux ont entrepris d'utiliser divers ensembles de données administratives pour estimer les flux migratoires. Le présent document traite de la production d'estimations des migrations intraprovinciales à l'aide de données sur les permis de conduire en Ontario. Une évaluation de ces estimations de la migration est faite à l'aide d'une comparaison des estimations produites à partir des données de l'impôt sur le revenu de Statistique Canada. Les deux fichiers donnent des estimations tout aussi bonnes et complémentaires des migrations intraprovinciales.

MOTS CLÉS: Fichiers administratifs; estimations de population; méthode des composantes; petites régions; erreur en fin de période; migration intraprovinciale.

1. INTRODUCTION

La migration est une composante importante des projections et des estimations de la population. Étant donné qu'on ne tient aucun fichier concernant le mouvement de la population au Canada, les démographes du gouvernement fédéral et ceux des administrations provinciales ont entrepris d'utiliser divers ensembles de données administratives pour estimer les flux migratoires. Statistique Canada utilise les données sur le revenu (Norris et Standish 1983), la Colombie-Britannique utilise les listes d'abonnés des services d'hydro-électricité (McRae 1985) et l'Alberta utilise les dossiers sur les bénéficiaires des soins de santé (Bureau de la statistique de l'Alberta 1985). Depuis 1979, l'Ontario utilise les changements d'adresse des détenteurs de permis de conduire pour estimer les migrations intraprovinciales. Outre leur fiabilité, les données sur les permis de conduire ont cet autre avantage important d'être différentes rapidement. En effet, le délai entre la date de réception des données et la date de référence n'est que de 4 à 5 semaines comparativement à plus d'une année et demie dans le cas des données sur le revenu. Dans le présent document, nous évaluons les estimations des migrations intraprovinciales produites à partir des données sur les permis de conduire en Ontario. Aux États-Unis, l'État de Californie utilise aussi les changements d'adresse des détenteurs de permis de conduire pour estimer les mouvements migratoires à l'intérieur de l'État (Hoag 1984).

2. FICHER DES DONNÉES SUR LES PERMIS DE CONDUIRE

L'information relative aux changements d'adresse des détenteurs de permis de conduire est fournie par le ministère des Transports et des Communications de l'Ontario. Un conducteur doit aviser le Ministère de sa nouvelle adresse dans les 90 jours suivant la date de son changement d'adresse. L'information est ventilée par région de code postal. Ces régions peuvent ensuite être converties en régions intraprovinciales, comme des comtés et des municipalités. Comme le tableau 1 l'indique, les données peuvent être obtenues pour les sept dernières années. Depuis 1979, les données peuvent également être produites par trimestre.

¹ Version abrégée d'un document présenté aux réunions du Comité fédéral-provincial de la démographie, les 28 et 29 novembre, 1985, Statistique Canada, Ottawa.
² Raghubar D. Sharma et Cheuk Wong, Direction des politiques sectorielles et régionales, ministère du Trésor et de l'Economie de l'Ontario, Queen's Park, Toronto (Ontario), M7A 1Y9.

- STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population des divisions et régions métropolitaines de recensement (méthode des composantes)*. Ottawa, N° 91-212 au catalogue, Ottawa, ministre des Approvisionnement et Services Canada.
- STOCK, R. (1981). *Migration Estimates from Current Administrative Files: Data Sources and Methodologies*. Canadian Plains Research Center, Université de Régina.
- U.S. BUREAU OF THE CENSUS (1973). *The Methods and Materials of Demography*. Washington, D.C.: U.S. Government Printing Office.
- VERMA, R.B.P., BASAVARAJAPPA, K.G., et BENDER, R.K. (1984). The Regression Estimates of Population for Sub-provincial Areas in Canada. *Techniques d'enquête*, 9, 219-240.
- VERMA, R.B.P., et BASAVARAJAPPA, K.G. (1985). Recent Developments in the Estimation of Population for Small Areas in Canada by Regression. Document présenté au Symposium international sur les statistiques régionales, Ottawa, Canada.

4.7 Statistiques du camionnage (entreprises de déménagement)

Il est possible d'obtenir les statistiques tirées d'un échantillon de cinq entreprises de déménagement parmi les principales au Canada. On pourrait évaluer les flux migratoires interprovinciaux en pondérant le nombre de déménagements déclarés entre deux provinces ou territoires. L'utilité des statistiques du camionnage est toutefois sérieusement diminuée par le fait qu'elles datent toujours d'au moins deux ans.

5. CONCLUSIONS

Dans le présent rapport, nous avons présenté une vue d'ensemble des avantages et inconvénients de douze fichiers de données administratives dans le but de faire des recommandations concernant le choix d'une source de données autre que les fichiers des allocations familiales. Nous avons constaté qu'aucun fichier n'avait des avantages et des inconvénients strictement comparables à ceux des fichiers des allocations familiales. Toutefois, advenant le cas où le programme des allocations familiales cesserait d'être universel, nous pourrions faire les recommandations suivantes:

- continuer d'explorer la possibilité d'utiliser le fichier de l'assurance-maladie des provinces et les dossiers de la sécurité de la vieillesse du gouvernement fédéral pour produire des estimations de la population totale et de la migration sur une base trimestrielle; - examiner la qualité des estimations annuelles de la population des provinces et des territoires produites par la Méthode des composantes II à l'aide des estimations de la migration générées à partir des fichiers provinciaux des données sur les effectifs scolaires; - vérifier la qualité des statistiques obtenues en utilisant les fichiers provinciaux de données administratives (fichiers de l'assurance-maladie, abonnés à l'Hydro, compagnies de téléphone et permis de conduire) comme indicateurs symptomatiques des variations de la population et de la migration nette résiduelle pour les régions intraprovinciales (divisions de recensement et régions métropolitaines de recensement au Canada).

BIBLIOGRAPHIE

ALMOND, M.M. (1982). An Inventory of Sources of Canadian Migration Data. Document de travail, Division de la démographie, Statistique Canada.

McRAE, D.G. (1985). Utilisation des comptes de l'hydro dans le modèle d'estimation par régression de la population en Colombie-Britannique. Document présenté au Comité fédéral-provincial de la démographie, Ottawa, Canada.

NORRIS, D.A., et STANDISH, L.D. (1983). Un Rapport technique sur la production d'estimations des migrations à partir de dossiers fiscaux. Rapport technique, Division de l'exploitation des données administratives, Statistique Canada.

NORRIS, D.A. (1983). Nouvelles sources de données sur la migration à l'échelle des petites régions. *Review of Public Data Use*, 11-25.

STATISTIQUE CANADA (Trimestriel). *Estimations trimestrielles de la population du Canada, des provinces et des territoires*. N° 91-001 au catalogue, Ottawa ministre des Approvisionnement et Services Canada.

STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population suivant l'état matrimonial, l'âge, le sexe et composantes de l'accroissement, Canada, provinces et territoires*. N° 91-210 au catalogue, Ottawa, ministre des Approvisionnement et Services Canada.

STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population des divisions et régions métropolitaines de recensement (méthode de régression)*. N° 91-211 au catalogue, Ottawa, ministre des Approvisionnement et Services Canada.

Le fait que la date à laquelle ces données s'appliquent est la date d'émission du permis. Pour cette raison, on ne peut être certain que la maison a effectivement été construite, ni qu'elle a été occupée. Un autre inconvénient du fichier, c'est que le nombre de permis émis n'est pas nécessairement directement lié aux variations de la population, surtout dans le cas d'une population décroissante.

Par conséquent, l'utilisation des données sur les permis de construire semble également présenter un potentiel limité pour estimer la population des diverses régions géographiques.

4.3 Commission d'assurance-chômage

La Commission d'assurance-chômage tient à jour une liste des bénéficiaires du programme. Un échantillon couvrant 10% de ce fichier a été prélevé pour produire des statistiques et pourrait produire des renseignements sur les migrations. Toutefois, ce fichier pourrait difficilement être utilisé pour estimer les migrations au Canada, et ce pour deux raisons. Premièrement, un échantillon couvrant 10% de l'univers des chômeurs correspond à moins de 1% de la population totale. Il est impossible de générer des statistiques sur les flux migratoires entre les provinces à partir d'une sous-population aussi petite. Deuxièmement, le fait que cet échantillon ne soit pas représentatif (les jeunes adultes représentant une bonne partie des chômeurs) invite à la prudence relativement aux données sur les migrations tirées de ce fichier. La Commission tient également à jour un fichier des salaires qui versent des cotisations d'assurance-chômage. Aucune analyse approfondie de ce fichier n'a toutefois encore été faite.

4.4 Enquête sur la population active

En 1982, Statistique Canada a réalisé une enquête-échantillon auprès de 56,000 ménages au Canada. On a posé aux enquêtés de la population civile hors institutions âgée de 15 ans et plus qui faisait partie de l'échantillon une question concernant leurs antécédents migratoires des cinq ou six dernières années. D'autres renseignements utiles peuvent également être obtenus. Toutefois, la très petite taille de l'échantillon (1/2% de la population totale) et le fait que l'enquête n'a été menée qu'une seule fois éliminent l'enquête sur la population active comme source de données à des fins d'estimation des migrations.

4.5 Listes électorales

Il est en général possible d'obtenir des données sur les électeurs au Canada. Les listes électorales fédérales et provinciales pourraient facilement être obtenues, tandis qu'il faudrait plus de travail pour obtenir les listes électorales municipales. Ces listes contiennent des renseignements sur le nombre de citoyens canadiens âgés de 18 ans et plus (les immigrants reçus sont inclus au niveau municipal seulement). Elles couvrent en moyenne 90 à 95% de la population cible. La principale lacune de cette source de données réside dans le fait que les statistiques ne peuvent être obtenues à des intervalles réguliers. Les listes électorales fédérales et provinciales sont dressées pour des élections qui ont lieu à peu près tous les quatre ans environ à des dates qui ne conviennent pas à des fins d'estimation. Il semble donc inutile d'envisager d'utiliser les listes électorales.

4.6 Ventes au détail

Des données sur les ventes au détail sont recueillies par Statistique Canada sous forme de chiffre d'affaires auprès de tous les grands magasins et d'un échantillon d'entreprises plus petites. Ces données sont recueillies mensuellement et peuvent être obtenues trois mois après la date de référence. L'utilité de cette série de données semble toutefois limitée sur le plan de l'estimation de la population et de la migration. Cela pourrait être attribuable au fait que les ventes au détail sont très fortement tributaires des fluctuations économiques qui peuvent ne pas toujours traduire fidèlement les variations dans la taille de la population.

mal (le 30 septembre au lieu du 1^{er} juin) et des délais pouvant atteindre dix mois sont d'autres handicaps. Au niveau infraprovincial, enfin, on constate souvent qu'un certain nombre d'étudiants résident dans une région administrative donnée, mais vont à l'école dans une autre. Cela aussi risque d'influer sur la qualité des estimations. Il faut remarquer ici que les données sur les effectifs scolaires ont déjà été utilisées par Statistique Canada (et aussi par le U.S. Bureau of the Census, à l'aide d'une méthode des composantes mise au point par ce bureau; voir U.S. Bureau of the Census 1973, Chap. 23, p. 51); les écarts obtenus à l'aide de cette dernière méthode ont été beaucoup plus grands que les écarts obtenus à l'aide d'autres méthodes. Cependant, si aucun autre fichier ne pouvait produire d'estimations adéquates de la population, les estimations obtenues par régression au moyen de ce fichier pourraient être acceptables, au moins au niveau provincial.

4. FICHIERS PRÉSENTANT UN POTENTIEL LIMITE

4.1 Permis de conduire

Toutes les provinces ont un fichier des personnes âgées de 15 (ou de 16 ou de 17) ans et plus qui ont un permis les autorisant à conduire un véhicule automobile. À l'aide de ces fichiers provinciaux, les migrations pourraient être estimées de deux façons: 1) par compilation des changements d'adresse sur les permis de conduire pour estimer les flux migratoires (et 2) en utilisant les fichiers comme indicateurs symptomatiques des variations de la population établies d'après les variations enregistrées dans le nombre des personnes ayant un permis de conduire dans une région donnée. À l'heure actuelle, l'Ontario utilise les permis de conduire pour estimer les migrations intraprovinciales, mais très peu d'autres provinces pourraient fournir de renseignements sur les flux migratoires, en particulier au niveau infraprovincial. La fourniture de tels renseignements aurait exigé trop de travaux et nécessité des consultations avec le ministre de la province. En dépit du fait que la loi oblige les conducteurs à indiquer leur changement d'adresse, tous ne le font pas et il n'y a pas de statistiques suffisamment détaillées.

Le fichier des permis de conduire pourrait également être utilisé dans les techniques de régression. La possibilité d'obtenir des données à n'importe quelle date précise et rapidement est un élément positif. La couverture de l'univers et les problèmes de cohérence pour- raient toutefois réduire la qualité des données. Par exemple, 83% des adultes en Saskatchewan ont un permis de conduire, comparativement à 73% au Manitoba (85% des hommes et 62% des femmes de cette province ont un permis de conduire). En outre, la proportion de ceux qui ont un permis de conduire chez les pauvres, les immigrants relativement récents, les Indiens et les résidents des collectivités éloignées du Nord est inférieure aux moyennes (Stock 1981, p. 44). Pour faire une estimation, il est souvent préférable d'avoir une couverture à 100% d'une petite partie de la population (par exemple les enfants) plutôt qu'une couverture à 80 ou 85% d'une grande partie de la population (par exemple les adultes), surtout si la couverture est sélective sur le plan de la migration. Même si cela ne fait pas nécessairement que le fichier ne convient pas pour l'estimation, cela en affecte le potentiel.

4.2 Permis de bâtir

Statistique Canada recueille des données sur les nouveaux permis de bâtir au niveau des villes et des régions rurales au Canada. En moyenne, les taux de couverture diffèrent selon qu'il s'agit de villes (98.5%) ou de régions rurales (62.5%). Il est possible d'obtenir des données mensuelles sur les permis de bâtir au niveau des divisions de recensement. Ces statistiques pourraient aussi être utilisées comme indicateurs symptomatiques des variations de la population. Un des inconvénients des données sur les permis de bâtir réside toutefois dans

d'hydro-électricité ou les abonnés résidentiels du téléphone pourraient être utilisés dans les techniques de régression comme indicateurs symptomatiques (voir McKae 1985, pour une application des données sur les abonnés des services d'hydro-électricité à l'estimation de la population). Cette méthode et les sources correspondantes sont généralement utilisées pour produire des estimations de la population dans les petites régions, mais si aucune autre méthode ne permet d'obtenir des estimations valables au niveau provincial, ces sources seront sérieusement prises en considération.

3.1 Abonnés des services d'hydro-électricité

Les compagnies d'électricité conservent des dossiers sur leurs clients. Ces dossiers renseignent sur le type de compte (résidentiel, commercial, exploitation agricole, etc.) et l'adresse et le code postal de l'abonné. Dans certaines provinces, il y a un seul fichier pour la province, mais, dans d'autres, deux compagnies (au Manitoba et à Terre-Neuve) ou même plus (en Colombie-Britannique et en Ontario) fournissent l'électricité à la population de la province. Par contre, dans la plupart des provinces, des statistiques peuvent être produites pour le territoire entier, à tout moment et rapidement, mais il y a quand même un petit nombre de provinces pour lesquelles il peut être difficile d'obtenir des données. Les principales faiblesses des fichiers sont de deux types. Outre le problème mentionné plus haut, il est possible qu'il y ait de légères incohérences à cause de la différence qui existe dans les définitions provinciales des ménages résidentiels (puisque'ils satisfont aux critères administratifs) et même à l'intérieur d'une même province lorsque plus d'une compagnie dessert le territoire. Néanmoins, les abonnés des services d'hydro-électricité pourraient être une très bonne source de données pour l'estimation de la population. En fait, cette source de données a été utilisée en Colombie-Britannique, où des estimations de la population ont été produites pour les municipalités et les districts scolaires, et les résultats ont été bons. Cette méthode pourrait également être utilisée et éventuellement élargie pour produire, au besoin, des estimations au niveau provincial.

3.2 Compagnies de téléphone

Au Canada, la majorité des services téléphoniques sont assurés par 14 principales compagnies de téléphone. Il est possible d'obtenir de l'information sur les abonnés des lignes résidentielles (adresse et code postal). La situation est sensiblement la même que celle des fichiers des abonnés à l'Hydro. Les données peuvent en général être obtenues assez rapidement et la couverture de l'univers est assez élevée. Encore ici, il est possible que plus d'une compagnie desserve un territoire donné et, aussi, qu'une même compagnie desserve plus d'une province. Malgré qu'aucune estimation fondée sur les fichiers des compagnies de téléphone n'ait encore été tentée par Statistique Canada, on estime que ces fichiers ont le potentiel pour produire de bons résultats.

3.3 Inscriptions scolaires

Chaque gouvernement provincial tient à jour un fichier informatisé sur les étudiants inscrits dans son réseau d'écoles primaires et secondaires; ce fichier contient des renseignements sur les adresses des écoles avec le code postal et le nombre d'étudiants, selon l'âge et le niveau d'études. L'information sur le nombre d'étudiants se rapporte au 30 septembre de chaque année et peut être obtenue de quatre à dix mois après la date de référence, les délais variant selon les provinces. La couverture de l'univers des étudiants est également très bonne. Ce fichier présente certains inconvénients. Par exemple, son caractère annuel empêche qu'il soit utilisé pour produire des estimations trimestrielles. Une date de référence convenant

les immigrants nouvellement arrivés et les étudiants étrangers) sont couverts par l'assurance-maladie de la province, à l'exception du personnel de la GRC et des Forces armées et des détenus des prisons fédérales qui sont, eux, couverts par l'assurance-maladie du gouvernement fédéral. Tous ceux qui établissent leur résidence dans une province donnée doivent remplir une demande en bonne et due forme, à partir de laquelle des données peuvent être compilées sur les immigrants internes par province d'origine et sur les immigrants provenant de l'étranger. La couverture presque complète de l'univers, la possibilité d'avoir des données mensuelles, les délais minimums et les données habituellement désagrégées selon l'âge, le sexe et la composition de la famille des migrants sont les principaux avantages de ces fichiers. Les migrants interprovinciaux devraient être fortement incités à faire une demande pour être couverts par ce programme. Par conséquent, les données sur les migrations provenant de ces fichiers devraient être fiables.

Ces fichiers présentent toutefois certains inconvénients. Leur principale lacune réside dans le fait que ni l'Ontario ni le Québec ne sont en mesure de fournir des données sur les migrations. Des changements s'annoncent dans le cas du Québec mais, pour l'Ontario, il n'y a rien à attendre. À moins de pouvoir générer une source spéciale pour l'Ontario, cela diminuerait le potentiel de ce fichier. Ce fichier risque également de poser des problèmes de cohérence étant donné que chaque province gère son fichier de façon indépendante. Au niveau infraprovincial, il serait également possible de générer des données sur les migrations étant donné que les bureaux des services de santé des provinces devraient être informés de tout changement d'adresse. Dans les faits, cependant, ce n'est pas toujours le cas. On pourrait également utiliser les fichiers des services de santé dans les estimations par régression, en particulier dans les provinces qui vérifient périodiquement les adresses pour épurer leur fichier et compter uniquement la population voulue.

2.2 Dossiers de la sécurité de la vieillesse

Santé et Bien-être social Canada est responsable de l'administration du fichier de la sécurité de la vieillesse. Les résidents canadiens âgés de 65 ans et plus qui totalisent suffisamment d'années de résidence au pays sont admissibles. Ils représentent environ 10% de la population totale. Presque toutes les personnes admissibles sont couvertes par le programme. En outre, l'incitation financière à signaler un changement d'adresse est très forte. Un autre avantage du fichier, c'est l'actualité de ses données. Les renseignements sur les personnes qui ont déménagé un mois donné seront compilés et reçus à Statistique Canada deux ou trois mois plus tard. Enfin, la sécurité de la vieillesse, étant un programme fédéral, fournit des données comparables d'une province à l'autre; même si les renseignements sont compilés par les bureaux provinciaux régionaux, la procédure est la même partout au Canada. La principale lacune de ce fichier, du point de vue l'estimation des migrations, réside dans le fait qu'il concerne une faible partie de la population (variant de 7,3% en Alberta à 12,2% à l'Île-du-Prince-Édouard) et, qui plus est, les personnes âgées ont un comportement migratoire assez différent de celui du reste de la population. Contrairement à la migration chez les enfants, qui peut être évidemment liée à la migration totale, le fichier de la sécurité de la vieillesse ne permet pas d'élaborer de méthode analogue et aussi efficace pour estimer la migration totale. Même si le fichier de la sécurité de la vieillesse n'a pu être utilisé comme source principale de données pour produire des estimations de la migration, il doit être considéré comme une estimation très intéressante de la migration chez les personnes âgées.

3. FICHIERS PRÉSENTANT LE PLUS DE POTENTIEL COMME INDICATEURS SYMPTOMATIQUES DES VARIATIONS DE LA POPULATION ET DES MIGRATIONS NETTES

Les données tirées de certains fichiers administratifs pourraient servir à générer des estimations de la population totale. Par exemple, les effectifs scolaires, les abonnés des services

Tableau 1 (suite)

Critères		Programme F55		Fichier de Revenu Canada	
Univers	Enfants ayant droit aux allocations familiales	Souscripteurs de déclaration d'impôt (doivent avoir rempli une déclaration d'impôt deux années consécutives)	Les souscripteurs dont les déclarations d'impôt s'appartiennent deux années consécutives représentent environ 75% de la population âgée de 18 ans et plus	Comparaison de l'adresse de retour figurant sur les déclarations d'impôt appariées. Des corrections sont apportées aux déclarations d'impôt non appariées	Intraprovinciale, Interprovinciale et Internationale
Coverture de l'univers	Analogue à la couverture des statistiques mensuelles des allocations familiales	Indicateur symptomatique (estimations des variations de la population totale à partir des variations de la population couverte)	Migrations nettes indirectes	Nombre d'enfants par région géographique. Âge	Deux fois par année, le 1 ^{er} juin et le 1 ^{er} décembre (désigne le nombre d'enfants ayant droit aux allocations familiales à ces dates)
Date ou période de référence	Deux fois par année, le 1 ^{er} juin et le 1 ^{er} décembre (désigne le nombre d'enfants ayant droit aux allocations familiales à ces dates)	Année: désigne la période entre la production d'une déclaration d'impôt deux années consécutives, c'est-à-dire à peu près la période d'avril d'une année à mars de l'année suivante. Les données relatives à cette période servent comme données de la période de juin à mai	Des données préliminaires peuvent être obtenues six à huit mois après la fin de la période de référence, alors que les données finales peuvent être obtenues dix à douze mois après la fin de la période de référence	De 1966-1967 à aujourd'hui	Des modifications aux lois fiscales ont entraîné des changements au niveau de la couverture et du nombre de déclarations d'impôt appariées dans le temps et selon les provinces
Possibilité d'avoir des séries chronologiques	De décembre 1977 jusqu'à aujourd'hui, pour l'information concernant les enfants ayant droit aux allocations familiales. Il est possible d'avoir des statistiques qui remontent à 1974 dans le cas des enfants pour lesquels des allocations familiales ont été versées				
Délais	Les statistiques peuvent être obtenues environ trois mois après la date de référence				
Caraçtéristiques	Nombre d'enfants par région géographique. Âge				
Types de migrations	Migrations nettes indirectes				

Description des fichiers de données administratives actuellement utilisés pour générer les données sur les migrations au Canada

Critères		Fichiers M0024 des allocations familiales	
Univers	Enfants pour lesquels une allocation familiale est versée	Enfants ayant droit aux allocations familiales (par opposition aux enfants pour lesquels une allocation familiale est versée)	Analogue à la couverture des statistiques mensuelles des allocations familiales
Couverture de l'univers	25% de la population totale en 1984. Presques 100% de tous les enfants âgés entre 0 et 17 ans.	statistiques mensuelles des allocations familiales	Analogue à la couverture des statistiques mensuelles des allocations familiales, en plus de la migration internationale
Méthode de détermination de la migration	Compilation des avis de changement d'adresse	Compilation des avis de changement d'adresse	Analogue aux types de migration couverts par les statistiques mensuelles des allocations familiales, en plus de la migration internationale
Types de migration	Migration interprovinciale, selon la province d'origine et de destination	Analogue aux types de migration couverts par les statistiques mensuelles des allocations familiales, en plus de la migration internationale	Origine-destination. Âge: total des enfants de 0 à 17 ans seulement. Taille de la famille: désigne le nombre d'enfants dans la famille
Caractéristiques	Origine-destination. Âge: total des enfants de 0 à 17 ans seulement. Taille de la famille: désigne le nombre d'enfants dans la famille	Mois: désigne le volume de renseignements traités durant cette période.	Données diffusées deux fois par année. Rensérme des informations sur la migration des six derniers mois. Les statistiques peuvent être obtenues trois mois environ après la fin de la version semi-annuelle. De décembre 1977 à aujourd'hui
Date ou période de référence	Mois: désigne le volume de renseignements traités durant cette période.	Données diffusées deux fois par année. Rensérme des informations sur la migration des six derniers mois. Les statistiques peuvent être obtenues trois mois environ après la fin de la version semi-annuelle. De décembre 1977 à aujourd'hui	Possibilité d'avoir des séries chronologiques
Délais	Les données traitées un mois donnée peuvent être obtenues à la fin de ce mois et concernent les migrations d'environ deux mois auparavant	À partir de janvier 1974 jusqu'à aujourd'hui pour les données concernant les migrations d'enfants. De 1947 à 1973, seules des données sur les migrations de familles peuvent être obtenues	Cohérence
Degré d'information	Dans les bureaux provinciaux, oui. Mais les données sont envoyées au bureau central de Santé et Bien-Etre social Canada sous forme d'imprimés	Dans les bureaux provinciaux, oui. Mais les données sont envoyées au bureau central de Santé et Bien-Etre social Canada sous forme d'imprimés	

Les douze fichiers administratifs sont évalués du point de vue qualitatif comme source de données pouvant remplacer les fichiers des allocations familiales. Ces fichiers ont été divisés en trois groupes principaux de la façon suivante en fonction de leurs avantages et de leurs inconvénients.

- Fichiers présentant le plus de potentiel pour l'estimation des flux migratoires*
- i) Fichiers de l'assurance-maladie
 - ii) Fichier de la sécurité de la vieillesse
- Fichiers présentant le plus de potentiel comme indicateurs symptomatiques des variations de la population et de la migration nette*
- iii) Abonnés des services d'hydro-électricité
 - iv) Abonnés du téléphone
 - v) Inscriptions scolaires
- Autres fichiers, présentant un potentiel limité ou incertain pour l'estimation des flux migratoires ou des migrations nettes*
- vi) Permis de conduire
 - vii) Permis de bâtir
 - viii) Bénéficiaires de l'assurance-chômage
 - ix) Enquête sur la population active
 - x) Listes électorales
 - xi) Ventes au détail
 - xii) Statistiques du camionnage

1.1 Critères d'évaluation des fichiers de données administratives

L'utilité des diverses sources de données administratives pour l'estimation des migrations interprovinciales et intraprovinciales est évaluée en fonction de dix critères: univers, couverture de l'univers, méthode de détermination de l'information sur les migrations, types de migration, caractéristiques des dossiers, date ou période de référence (et possibilité d'avoir des données mensuelles), délai, possibilité d'obtenir des séries chronologiques, cohérence et informatisation (Almond 1982).

La nouvelle source de données à utiliser présenterait un grand potentiel si elle contenait les caractéristiques des fichiers des allocations familiales, décrites au tableau 1. Les critères les plus importants sont les suivants: couverture, actualité des données, cohérence, possibilité d'obtenir des données mensuelles ou trimestrielles, niveau de désagrégation utilisant le code postal ou d'autres systèmes de géocodage. Le fichier ou l'ensemble de fichiers qui pourrait satisfaire ces critères aurait probablement les qualités voulues pour remplacer les fichiers des allocations familiales.

2. FICHIERS PRÉSENTANT UN GRAND POTENTIEL POUR L'ESTIMATION DES FLUX MIGRATOIRES

Les fichiers de l'assurance-maladie et de la sécurité de la vieillesse sont ceux qui ont le plus de potentiel pour l'estimation des flux migratoires entre les provinces, territoriales et divisions de recensement. Les avantages et les inconvénients de chacun de ces fichiers sont décrits ci-après.

2.1 Fichier de l'assurance-maladie

L'assurance-maladie est du ressort des provinces. Chaque province tient donc à jour un dossier des personnes admissibles au programme. Tous les résidents d'une province (y compris

Vue d'ensemble des avantages et inconvénients des fichiers de données administratives¹

RAVI B.P. VERMA et PIERRE PARENT²

RÉSUMÉ

Dans ce texte, on évalue les possibilités de tirer des données de migration de douze fichiers de données administratives, dans l'éventualité où l'universalité des données d'allocation familiales, qui sont utilisées présentement, serait remise en question par une loi fédérale. Il est montré qu'aucun fichier ne présente des avantages et des inconvénients exactement comparables à ceux de l'allocation familiale. Il est fortement recommandé de poursuivre le développement des fichiers d'assurance-santé, et, dans une moindre mesure du fichier de la sécurité de la vieillesse.

MOTS CLÉS: Fichiers administratifs; migration; évaluation qualitative.

1. INTRODUCTION

Au Canada, les fichiers des allocations familiales et de l'impôt sur le revenu sont utilisés de diverses façons pour produire des estimations de la migration et de la population pour les différentes régions géographiques (voir les publications nos 91-001, 91-210, 91-211 et 91-212 au catalogue). Les données provenant des fichiers des allocations familiales sont publiées deux à trois mois après la date de référence, tandis que les données tirées des fichiers de l'impôt sur le revenu sont, elles, publiées douze à quinze mois après la date de référence. Cependant, les données de l'impôt sur le revenu produisent des estimations des flux migratoires par division de recensement et également selon l'âge et le sexe.

En termes de précision des estimations de la population, les deux types de fichiers donnent de bons résultats et sont comparables (voir Norris et Standish 1983; Verma et al. 1984; Verma et Basavarajappa 1985). Une des caractéristiques particulières des fichiers des allocations familiales et des fichiers de l'impôt sur le revenu réside dans le fait qu'ils ont un caractère national. Autre caractéristique, les dossiers contiennent des adresses où figure le code postal, de sorte qu'il est possible d'obtenir les renseignements sur les migrations locales. Toutefois, depuis quelques années, il semble y avoir possibilité que les allocations familiales cessent d'être universelles par suite de mesures législatives gouvernementales. Par exemple, les secteurs couverts pourraient être limités aux secteurs de la population des tranches de revenu inférieures et intermédiaires. Si ce fichier cessait d'être universel, son utilité comme source de données sur les migrations pourrait être sérieusement réduite. Par conséquent, notre capacité d'établir des estimations de la population pourrait être affaiblie, ce qui pourrait, par voie de conséquence, influer sur d'autres programmes comme le programme ayant trait au partage des revenus, qui suppose la répartition annuelle de 20 milliards de dollars entre les provinces.

Pour cette raison, il faut explorer d'autres sources de données. Nous tentons ici d'évaluer les avantages et les inconvénients de certains des fichiers de données administratives choisis pour l'estimation des migrations et de la population dans les provinces et territoires, les divisions de recensement, les régions métropolitaines de recensement ainsi que dans d'autres régions du Canada.

¹ Version abrégée du document présenté à la réunion du Comité fédéral-provincial sur la démographie tenues les 28 et 29 novembre 1985 à Ottawa.

² Ravi B.P. Verma et Pierre Parent, Division de la démographie, Direction de la statistique démographique et du recensement, Statistique Canada, 4^{ème} étage, Immeuble Jean Talon, Tunney's Pasture, Ottawa (Ontario), Canada, KIA 0T6.

- STATISTIQUE CANADA (1979). Basic Questionnaire Design, deuxième édition. Document non publié, Statistique Canada.
- STATISTIQUE CANADA (1981). Conception des questionnaires, Manuel d'atelier, troisième édition. Document non publié, Statistique Canada.
- WARWICK, D.P., et LININGER, C.A. (1975). *The Sample Survey: Theory and Practice*. New York: McGraw-Hill.

La méthode de collecte utilisée pour l'essai préliminaire doit être identique à celle prévue pour l'enquête principale. Mais une interview sur place est recommandée pour au moins une partie des enquêtes de l'essai préliminaire afin que l'interviewer puisse noter les réactions, verbales ou autres, des répondants ainsi que leurs suggestions et impressions. À la fin de chaque interview de l'essai préliminaire, l'interviewer peut discuter des difficultés éprouvées par le répondant, de l'interprétation des questions et des catégories de réponse et ainsi de suite. Le concepteur du questionnaire peut ensuite organiser une réunion avec les interviewers qui ont participé à l'essai préliminaire pour discuter de ces difficultés et faire le bilan de l'expérience. Dans certains cas, il est préférable de faire appel à des interviewers qualifiés et expérimentés afin de tirer profit au maximum de l'essai préliminaire.

L'essai préliminaire est une étape souvent négligée. Il révèle presque toujours des améliorations qui peuvent être apportées ou permet du moins au concepteur du questionnaire de s'assurer que l'utilisation du questionnaire dans l'enquête principale, qui coûte beaucoup plus qu'un essai préliminaire, sera probablement assez efficace. Bien entendu, rien ne garantit jamais que tous les problèmes seront résolus, mais la plupart des grandes difficultés devraient l'être. Un essai préliminaire n'est pas nécessairement coûteux, ne prend pas forcément beaucoup de temps et est recommandé pour les questionnaires nouveaux ou modifiés.

REMERCIEMENTS

L'auteur remercie l'arbitre et le conseil de rédaction pour leurs observations utiles, ainsi que les nombreuses personnes qui ont examiné une version antérieure de ce texte qui faisait partie d'une ébauche partielle d'un document intitulé "Survey Design Standards and Guidelines".

BIBLIOGRAPHIE

- ANDERSON, J.F., et BERDIE, D.R. (1974). *Questionnaires: Design and Use*. Metuchen, N.J.: The Scarecrow Press, Inc.
- BERTHIER, N., et F. (1971). *Le sondage d'opinion*. Paris: Bordas.
- BON, F. (1974). *Les sondages - peuvent-ils se tromper?* France: Calmann-Lévy.
- CARSON, E. (1974). *Questionnaire Design, Some Principles and Related Topics*. Document non publié, Statistique Canada.
- CORBIN, R., SWAIN, L., et WILHELM, E. (1977). Exposé pour un atelier sur la conception des questionnaires. Document non publié, Statistique Canada.
- CORBIN, R., SWAIN, L., et WILHELM, E. (1977). Outline for a Workshop on Basic Questionnaire Design. Document non publié, Statistique Canada.
- GHIGLIONE, R., et MATALON, B. (1978). *Les enquêtes sociologiques: théories et pratique*. Paris: Colin.
- JAVEAU, C. (1974). *L'enquête par questionnaire. Manuel à l'usage du praticien*, troisième édition. Bruxelles: Institut de sociologie de l'Université libre de Bruxelles.
- MOSE, C.A., et KALTON, G.J. (1972). *Survey Methods in Social Investigation*, deuxième édition. New York: Basic Books.
- OPPENHEIM, A.N. (1966). *Questionnaire Design and Attitude Measurement*. Londres: Heinemann.
- PAYNE, S.L. (1951). *The Art of Asking Questions*. Princeton: Princeton University Press.

Dans les enquêtes sur les attitudes, le concepteur d'un questionnaire doit éviter de condonner au départ les répondants en fonction d'un schéma de référence qui peut fausser les réponses aux questions posées par la suite. Par exemple, les questions mesurant le degré de connaissance d'un sujet doivent précéder toutes les autres questions qui abordent ce sujet. Les questions délicates doivent être regroupées avec d'autres questions connexes pour justifier leur inclusion autant que possible et atténuer quelque peu l'embaras que le répondant peut éprouver.

13. La version finale du questionnaire ne doit contenir aucune erreur typographique que ou grammaticale.

La présence d'erreurs dans le questionnaire peut avoir un effet nuisible sur la qualité des données à cause du risque que le questionnaire ne soit pas pris au sérieux ou soit mal compris par les personnes qui le remplissent. Par ailleurs, les erreurs peuvent se répercuter négativement sur la manière dont le public perçoit l'organisme qui mène l'enquête.

5. ESSAIS

1. Les questionnaires qui sont utilisés pour la première fois ou ont été modifiés en profondeur doivent être mis à l'essai avant d'être employés pour la collecte de données.

Même si tous les principes décrits dans les élaborations précédentes ont été respectés et que le concepteur du questionnaire a été très consciencieux, il n'existe aucune garantie que le questionnaire proposé répondra entièrement aux objectifs de l'enquête. Il se produit presque toujours des problèmes imprévus lors de l'utilisation d'un questionnaire. Il est donc essentiel de faire un essai préliminaire du questionnaire pour toute enquête nouvelle ou très modifiée, afin de déterminer si le questionnaire proposé peut répondre aux objectifs de la recherche.

Le concepteur du questionnaire peut, par exemple, étudier la formulation, l'agencement et la disposition des questions pour voir si les répondants et les interviewers comprennent les questions et leur enchaînement; la nécessité d'inclure des questions particulières; le choix des genres de questions; l'emploi de questions spéciales comme celles où l'enquête doit classer un choix de réponses selon le rang ou appliquer une échelle d'évaluation; la structure et la définition des catégories de réponses; la fréquence de la catégorie "autre" dans les réponses; la facilité de remplir le questionnaire; le temps requis pour remplir les diverses sections du questionnaire; la traduction du questionnaire; la possibilité d'un biais dans les questions; la nature des différences ethniques, régionales ou linguistiques; le fardeau que le questionnaire impose au répondant; l'efficacité avec laquelle le questionnaire mesure les notions visées; les lettres ou autres moyens utilisés pour présenter l'enquête et la qualité de la méthode de collecte des données.

Il faut effectuer un essai préliminaire auprès d'au moins un petit échantillon de répondants (généralement une vingtaine ou une trentaine) de la population cible. Il est préférable de choisir ces répondants dans les divers sous-groupes de la population susceptibles de présenter des différences ou des problèmes. Les sous-populations pour un essai peuvent être définies en fonction de variables comme la région géographique, le niveau d'instruction, l'âge, le sexe, la langue, la taille de l'entreprise et la catégorie d'activité économique. Selon les buts particuliers de l'essai préliminaire, la sélection des répondants peut reposer sur un échantillon aléatoire ou non aléatoire mais, dans la plupart des cas, la méthode de sélection est non aléatoire. Il est également possible d'organiser un groupe de travail pour discuter du questionnaire dans le cadre de l'essai préliminaire.

Si un questionnaire doit être recueilli par un représentant sur le terrain, on doit prévoir un espace sur le questionnaire pour le nom, le numéro de téléphone ou l'adresse postale où il est possible de joindre le représentant et indiquer la date et l'heure approximative à laquelle ce dernier viendra.

10. Les chiffres et les codes utilisés pour la saisie des données doivent figurer sur le questionnaire (quand la saisie doit se faire directement à partir du questionnaire).

Dans certains cas, les données peuvent être saisies plus rapidement et avec moins d'erreurs directement à partir du questionnaire. Il faut alors que les chiffres et les codes puissent être lus facilement par les personnes chargées de la saisie des données, mais ils ne doivent pas nuire à la concentration du répondant, de l'interviewer ou d'un autre représentant qui remplit le questionnaire.

Quand les données doivent être codées avant la saisie, les cases pour le codage peuvent être imprimées sur le questionnaire ou sur une feuille séparée. Si les cases de codage figurent sur le questionnaire, elles doivent être clairement distinguées des cases pour les réponses, peut-être par une indication comme *réserve au bureau* ou par une couleur appropriée. Le codage et la saisie des données sont souvent considérés comme des étapes liées à la structure du questionnaire. Il est essentiel d'y penser pendant l'élaboration du questionnaire pour assurer l'efficacité de leur déroulement.

11. La disposition des espaces pour les réponses doit être uniforme dans tout le questionnaire et ces espaces doivent être assez grands pour que le questionnaire soit lisible et offre assez de place pour les réponses.

L'uniformité des espaces pour les réponses facilite le travail du répondant, de l'interviewer ou du représentant qui remplit le questionnaire et aide à réduire les erreurs dues à l'omission accidentelle d'une question, au choix d'une réponse inexacte ou à la transposition de réponses. Il peut être utile d'employer des formes d'espace différentes pour les questions fermées et celles dont la réponse est un chiffre. Des cercles sont parfois utilisés pour les questions fermées, des cases carrées pour les réponses numériques.

L'espacement doit être étendu pour plusieurs raisons: pour rendre le questionnaire facile à remplir, lui donner une bonne apparence, le rendre plus lisible et fournir assez de place au répondant, à l'interviewer ou au représentant pour inscrire la réponse à chaque question. 12. Les questions doivent se suivre dans un ordre logique pour en faciliter la lecture et établir le schéma de référence nécessaire.

L'ordre des questions doit sembler logique au répondant (dont la logique peut être différente de celle du concepteur du questionnaire) et les questions qui sont liées entre elles doivent se trouver ensemble. Une méthode qui est parfois recommandée consiste à commencer par les questions les plus générales avant de poser les questions les plus détaillées. L'ordre des questions doit autant que possible correspondre à l'ordre dans lequel les répondants fourniront des renseignements. Le concepteur d'un questionnaire doit se rappeler qu'une question peut susciter une réponse visant également une autre question qui (si l'ordre des questions est logique) n'est pas très loin. La transition entre les différentes sections du questionnaire ne doit pas être brusque. Un en-tête ou une explication doit figurer au début de chaque section. Si un questionnaire est utilisé pour transcrire des renseignements figurant sur d'autres documents, un ordre logique serait celui des documents dans lesquels les informations sont puisées.

5. Les répondants doivent être identifiés d'une manière convenable sur le questionnaire.

Pour l'estimation, le contrôle sur le terrain, l'enchaînement avec d'autres enregistrements ou les suivis auprès des non-répondants, les répondants doivent être identifiés (par un code numérique ou autrement) sur les questionnaires.

6. Les questions et les pages du questionnaire doivent être numérotées.

Pour faciliter la lecture du questionnaire pour les interviewers et les répondants et simplifier les opérations de codage et les instructions, les questions et les pages doivent être numérotées consécutivement (à l'aide de lettres ou de chiffres) dans l'ensemble du questionnaire. Si des questions sont imprimées sur les deux côtés d'une feuille, une indication (comme *suite au verso*) doit apparaître au bas du recto pour assurer qu'une réponse est fournie aux questions qui figurent sur le deuxième côté.

7. Les caractères d'impression doivent être faciles à lire pour le répondant moyen.

On doit penser à la personne qui répondra au questionnaire quand on décide de la grandeur des caractères d'impression (par exemple, de petits caractères pourraient poser des problèmes pour les personnes dont la vue n'est pas bonne) et qu'on choisit les couleurs et le degré de contraste entre le papier et les caractères. Il est normalement préférable d'utiliser des caractères différents (de grandeur ou de forme différente) pour les questions et les instructions, de sorte qu'on puisse facilement distinguer ces parties.

8. Des instructions concernant la manière de répondre aux questions doivent être imprimées sur le questionnaire ou l'accompagnager.

Pour aider les répondants, les interviewers ou d'autres représentants à bien remplir le questionnaire, des instructions brèves mais claires doivent être imprimées sur le questionnaire. Toutefois, les questions doivent être aussi claires que possible pour éviter d'avoir à donner des instructions compliquées.

Dans le cas des questionnaires lus par un lecteur optique de caractères, des instructions claires doivent être fournies pour assurer qu'ils sont bien remplis. Les instructions indiquant aux répondants ou aux interviewers les questions à omettre après les questions filtres doivent être assez simples et faciles à suivre. Il peut être approprié d'utiliser des flèches et des indications. On doit éviter des séries complexes de directives sur les questions à omettre, surtout dans les questionnaires remplis par les répondants eux-mêmes.

9. Les instructions pour le renvoi du questionnaire doivent figurer sur le questionnaire.

Si un questionnaire doit être retourné par la poste, le nom et l'adresse de la personne (ou de l'organisme) à laquelle il faut l'envoyer doivent être indiqués sur le questionnaire. Les lettres de présentation et les enveloppes de retour peuvent facilement se perdre ou être séparées du questionnaire. La date limite avant laquelle les répondants doivent retourner les questionnaires remplis doit également être précisée.

Dans les contextes où plusieurs mots peuvent être utilisés indifféremment, il faut choisir un mot et l'employer partout dans le questionnaire. Si un synonyme d'un mot qui figure ailleurs le remplace dans un autre passage, les répondants risquent d'y accorder un sens différent.

11. Les questions doubles doivent être évitées.

Une question double demande au répondant de fournir une seule réponse à ce qui est en réalité deux questions. La réponse ne permet pas de savoir à laquelle des deux questions le répondant pensait ou si ce dernier pensait aux deux à la fois. Les deux renseignements recherchés doivent être demandés séparément, sauf dans les cas particuliers où il faut nécessairement les demander en même temps pour transmettre le message voulu. On doit alors indiquer clairement aux répondants que les deux aspects de la question doivent être envisagés ensemble.

12. Les questions tendancieuses doivent être évitées.

Une question tendancieuse est formulée ou présentée de manière à porter le répondant à choisir une réponse ou série de réponses en particulier.

Certaines questions peuvent être considérées comme tendancieuses si elles renforcent des choix que les répondants peuvent trouver socialement inacceptables si on ne les assure pas que leur réponse ne suscitera aucun jugement de valeur.

Dans les enquêtes sur les attitudes, deux principes de base ont été établis pour réduire (mais pas nécessairement éliminer) le biais de réponse. Le choix des réponses doit être équilibré, de sorte qu'il y a autant de réponses positives que de réponses négatives pour éviter de faire pencher le répondant pour un type de réponse. Deuxièmement, s'il y a une série de questions qui ont le même choix de réponses, cette série doit soit contenir un mélange de phrases positives et négatives, soit être divisée, soit être présentée dans un ordre varié afin de réduire la possibilité que les répondants puissent répondre de la même manière pour toute la série (même si cela peut n'être pas approprié) sans penser très sérieusement à chaque réponse.

4. PRÉSENTATION

1. Chaque questionnaire ou ensemble de formules d'enquête doit contenir un texte qui explique et présente l'enquête.

2. Le texte d'introduction doit préciser le titre de l'enquête, le nom des organismes qui l'ont commandée et les objectifs visés.

3. Il faut songer à inclure un avis assurant les répondants quelles données fournies sont confidentielles.

4. Le questionnaire doit indiquer le nom (s'il y a lieu) et le numéro de téléphone ou l'adresse postale d'une personne avec qui les répondants peuvent entrer en contact dans l'organisme qui a commandé l'enquête s'ils ont besoin de renseignements supplémentaires.

En général, le texte de présentation peut être une lettre ou une brochure envoyée au répondant; ce texte peut également être rédigé à l'avance et lu par un intervieweur; ou il peut être imprimé sur le questionnaire. Ce texte doit fournir des renseignements généraux essentiels pour identifier l'organisme qui mène l'enquête, justifier l'enquête et informer les répondants de leurs droits (s'il y a lieu).

Le choix du type de question et de la formulation impose des restrictions sur la manière dont le renseignement demandé peut être utilisé ou analysé et les hypothèses précises qui peuvent être vérifiées. Il s'ensuit que la définition des objectifs, des utilisations, des hypothèses à vérifier et des analyses voulues doit se faire avant que ne soit fixée la forme définitive d'une question. Le fait que ces éléments soient établis à l'avance n'exclut pas la possibilité que les données recueillies puissent servir à d'autres analyses et besoins à l'intérieur des limites imposées par la nature des questions.

Outre les facteurs mentionnés plus haut, l'expérience du concepteur d'un questionnaire influera sur le choix des formes de questions qui conviennent à différentes situations.

7. Il ne doit pas y avoir de chevauchement dans les réponses aux questions fermées et ces réponses doivent être exhaustives (c'est-à-dire qu'elles sont mutuellement exclusives et complètes).

Les réponses à une question particulière doivent être distinctes et inclure toutes les possibilités.

Le fait que les réponses à une question fermée doivent être exclusives ne signifie nullement qu'aucune question ne peut avoir plus d'une réponse. S'il y a plus d'une réponse, une note telle que *cocher autant de réponses que nécessaire* doit être incluse avec la question. Si seulement une réponse doit être fournie à une question, une note telle que *cocher une case seulement* doit être incluse dans la question (sauf dans les cas les plus évidents, par exemple quand la réponse est *oui* ou *non*). Si plus d'une réponse peut être bonne mais que le concepteur veut que le répondant choisisse seulement une catégorie afin de limiter les réponses, une note telle que *cocher la réponse la plus appropriée* doit figurer dans la question.

8. Les unités de mesure de chaque réponse doivent être précisées.

Il faut soit préciser l'unité de mesure de la réponse (kilogrammes, tonnes, pourcentage, heures par semaine, etc.) à l'intérieur de la question, soit demander au répondant de l'indiquer lui-même. Autrement, il peut y avoir de la confusion au sujet de l'unité dans laquelle une réponse est exprimée.

9. Des notions et des définitions normalisées doivent être utilisées.

Pour faciliter la comparaison des données d'une enquête à celles d'autres sources d'information (publications, enquêtes) et rendre les données aussi utiles que possible (y compris pour les analyses secondaires), il est important d'employer des définitions normalisées (généralement comprises et utilisées) quand elles existent et qu'elles sont claires, appropriées et à jour. Statistique Canada publie des classifications types pour les professions, les activités économiques, les marchandises et les unités géographiques, de même qu'un répertoire de concepts sociaux. Par ailleurs, les notions et les catégories établies pour le recensement sont souvent utilisées comme normes.

10. La formulation des questions doit être précise, complète, cohérente, brève, simple et claire.

Les notions et les termes que les répondants ne connaissent pas ou peuvent mal comprendre doivent être expliqués, définis ou évités. Pour que les résultats soient interprétés de manière cohérente, le schéma de référence approprié (par exemple la période de référence, l'endroit visé ou le type de dépense étudié) doit être fourni. Si la cohérence des données risque d'être imparfaite (par exemple si différentes questions portent sur différentes périodes), tous les changements de schéma de référence doivent être soulignés dans le questionnaire.

3. Le concepteur d'un questionnaire doit choisir les types de question qui conviennent le plus aux renseignements demandés et permettent de réduire au minimum l'erreur de réponse et le fardeau des répondants.

Les principaux types de question sont les questions ouvertes à réponse libre, les questions fermées à réponses fixes et les questions à réponse structurée. Les questions fermées comprennent une liste des réponses possibles. Les questions à réponse structurée semblent ouvertes parce qu'aucun choix de réponses n'est fourni, mais elles sont implicitement fermées parce que le répondant est normalement obligé de se limiter à un chiffre, un jour de la semaine, une province, etc.

En général, le fardeau du répondant et/ou de l'interviewer est moins grand quand les questions sont fermées parce que les répondants n'ont pas à chercher leurs propres mots pour formuler des réponses et que ces réponses n'ont pas à être transcrites textuellement.

4. Lorsqu'un répondant doit choisir entre deux réponses bien définies ou plus, une question fermée ou à réponse structurée doit être utilisée.

Si les réponses à une question sont trop nombreuses pour être toutes énumérées, il est recommandé d'utiliser la catégorie *autre* pour englober des réponses peu fréquentes, de poser une question à réponse structurée ou à réponse libre ou de regrouper certaines réponses pour réduire le nombre de catégories. On peut même utiliser une question à réponse structurée si les réponses et leurs codes numériques figurent dans un livret d'instructions accompagnant le questionnaire. Dans les enquêtes sur les entreprises, l'agriculture et les institutions, des questions à réponse structurée sont souvent employées quand les réponses sont numériques. Le choix et le nombre de réponses à une question fermée dépend de la complexité de la notion mesurée, de la manière dont les données seront utilisées et des informations préalables dont le concepteur d'un questionnaire dispose.

5. Si le choix de réponses à une question n'est pas bien défini, une question ouverte doit être utilisée.

Les questions ouvertes sont souvent utilisées dans les recherches préliminaires ou exploratoires entreprises en vue de formuler des hypothèses précises et de structurer les questions pour les versions ultérieures d'un questionnaire. Les questions ouvertes offrent également un moyen d'obtenir des renseignements supplémentaires ou des précisions, de vérifier les réponses à d'autres questions, d'interpréter les données, de modifier le rythme du questionnaire ou d'aborder un nouveau sujet.

6. Si la facilité, la rapidité et le coût du traitement des données pour la saisie sont des facteurs importants, des questions fermées doivent être utilisées.

Les réponses à des questions ouvertes doivent être codées, opération qui peut être coûteuse, prendre beaucoup de temps et introduire des erreurs d'interprétation et de codage. Un autre problème est que les questions ouvertes ne fournissent aucun schéma de référence et que les répondants choisissent alors différents schémas pour leurs réponses. Comme les répondants peuvent avoir différents schémas de référence et fournir différentes quantités d'information, il est difficile d'interpréter, de coder et d'analyser les réponses. En revanche, une question fermée contient un schéma de référence précis qui permet d'éviter le problème susmentionné mais peut susciter une réponse artificiellement. Ce problème est particulièrement grave quand le répondant connaît peu un sujet donné, l'ignore ou n'a pas d'opinion marquée. Le concepteur du questionnaire doit donc considérer les schémas de référence possibles des répondants avant de décider de la forme d'une question.

3. Seules des questions auxqueltes il est possible de répondre facilement et avec assez de fiabilité doivent être incluses.

Quand les répondants doivent chercher dans leur mémoire, les événements visés doivent être assez récents ou connus des répondants; si les informations voulues se trouvent dans des dossiers des répondants, l'effort (mesuré en temps et en argent) nécessaire pour obtenir ces renseignements ne doit pas dépasser les avantages qui découleront de cette recherche. À cause du risque d'ambiguïtés dans les définitions, d'accroissement du fardeau des répondants et d'erreurs de traitement, il peut être préférable de ne pas demander aux répondants de traiter les données qui conduisent à un certain résultat. On peut parfois faciliter la collecte des données et éviter des erreurs en demandant aux répondants de fournir des informations dont ils disposent et laisser l'organisme qui mène l'enquête faire le traitement de ces renseignements par la suite.

4. Il ne faut pas poser de questions auxqueltes il n'est pas évident que les répondants peuvent répondre.

Les questions ne doivent pas présupposer que le répondant possède des connaissances sur un sujet précis ou exerce une activité particulière. Des questions filtres permettent d'omettre les questions qui ne s'appliquent pas à un répondant à cause de ses caractéristiques, de sa situation ou de ses opinions. Si les répondants trouvent que beaucoup de questions ne s'adressent pas à eux, ils risquent de penser que le questionnaire leur a été donné par erreur, ce qui peut aggraver la non-réponse ou nuire aux relations avec les répondants. Une question filtre indique clairement si un répondant doit répondre à une question ou à une série de questions. Cette propriété est utile pour le traitement et l'analyse des données d'une enquête. Si aucune réponse n'est fournie à une question, il peut être difficile de distinguer entre les cas de non-réponse (refus ou oubli accidentel) et les cas où une question n'est pas applicable (pour une réponse numérique, on risque de conclure que la question ne s'applique pas au répondant, alors que la réponse est en fait zéro). Une question filtre aide à résoudre ce problème en montrant quels répondants devaient répondre à une question. On doit toutefois éviter de rédiger des directives compliquées sur les questions à omettre, en particulier pour les questionnaires que les répondants remplissent eux-mêmes. Il faut aussi réduire au minimum le nombre de questions filtres. Pour les réponses numériques, on peut éviter de poser une question filtre en incluant la catégorie *aucun ou aucune*.

3. FORMULATION

1. La formulation d'une question doit convenir au répondant.

Si un répondant ne comprend pas une question, il donnera probablement une réponse inexacte ou ne répondra pas. Il faut que les mots, le style et la structure des phrases soient faciles à comprendre et conviennent aux personnes qui doivent fournir des renseignements. Il faut éviter les abréviations à moins que les répondants ne les comprennent.

2. Lorsque la demande est suffisante, un questionnaire doit être traduit en d'autres langues.

Il est important de vérifier si la traduction reproduit bien le sens du texte de départ.

sont énoncés, souvent ils peuvent ne pas l'être en pratique. De même, certains des principes peuvent être mesurés, alors que d'autres ne peuvent pas l'être.

Dans ce qui suit, le terme *questionnaire* représente toujours les diverses formules utilisées pour recueillir des informations. Dans la documentation et en pratique, on distingue souvent les documents suivants:

- a) questionnaires remplis par un répondant;
- b) questionnaires remplis par un intervieweur;
- c) formules administratives remplies par un répondant ou un représentant officiel de l'organisme qui mène une enquête;
- d) formules utilisées pour enregistrer des observations ou des mesures (remplies par un représentant officiel de l'organisme qui mène une enquête);
- e) formules utilisées pour transcrire des informations figurant dans des dossiers administratifs (remplies par un représentant officiel de l'organisme qui mène une enquête).

Pour simplifier les choses, on emploie ici le terme *questionnaire* de manière à englober tous ces sens. En outre, on entend par *question* chaque question ou demande d'information, y compris le choix de réponses ou l'espace prévu pour une réponse.

Le terme *enquête* est utilisé dans un sens général pour toute activité de collecte de données comme les enquêtes par sondage, les recensements et le prélèvement de données administratives.

2. CONTENU

1. Toutes les questions d'un questionnaire doivent porter directement sur l'objectif et les fonctions de l'enquête.

Il est raisonnable d'exiger que la collecte des données soit organisée de manière à réduire au minimum le fardeau des répondants. Un moyen de satisfaire à cette condition est d'exclure les questions qui n'ont qu'un rapport vague avec les objectifs et les fonctions de l'enquête. Les demandes de renseignements inutiles augmentent aussi la longueur d'un questionnaire et peuvent provoquer une attitude de méfiance chez le répondant, facteurs susceptibles d'accroître les taux de non-réponse (source possible d'erreurs systématiques), de baisser la qualité des données en fatiguant les interviewers ou les répondants ou en diminuant leur concentration et de coûter plus d'argent et de temps au client qui a commandé une enquête et aux répondants.

Du point de vue du concepteur d'un questionnaire, le fait même de préciser le lien entre chaque question et les objectifs et les fonctions de l'enquête l'aide à s'assurer que ces objectifs et fonctions sont bien définies et que le questionnaire y correspond effectivement.

2. Si un questionnaire demande des renseignements qui sont utiles pour l'enquête mais risquent de ne pas le sembler pour les répondants, il faut alors expliquer aux répondants la raison pour laquelle ces informations sont nécessaires.

Des variables de classification comme l'âge, le sexe, l'état matrimonial, la taille d'un organisme, l'effectif d'un employeur et des variables telles que le nom, l'adresse et le numéro de téléphone (renseignements utilisés pour des suivis ou des vérifications) sont des exemples possibles de données dont l'enquêteur doit songer à expliquer la nécessité (du moins à un niveau général) aux répondants.

Principes fondamentaux pour le développement des questionnaires

LARRY SWAIN¹

RÉSUMÉ

Trente principes fondamentaux sont présentés à l'égard du développement des questionnaires comprenant le contenu, la formulation, la présentation et l'évaluation des questionnaires. L'importance du questionnaire comme composant intégral d'une enquête est soulignée comme aussi la considération de son rapport avec d'autres aspects du plan de l'enquête.

MOTS CLÉS: enquête; questionnaire; méthodologie.

1. INTRODUCTION

La plupart des enquêtes reposent sur un questionnaire qui doit être rempli par le répondant ou un représentant officiel de l'organisme qui mène l'enquête (sur place ou par téléphone). Comme le questionnaire est l'instrument qui traduit les objectifs d'une enquête en variables mesurables, la réalisation de ces objectifs exige un questionnaire efficace. En outre, le questionnaire peut aider à organiser, à normaliser et à contrôler la collecte des données de telle sorte que les informations nécessaires puissent être recueillies d'une manière satisfaisante. Pour bien élaborer un questionnaire, il faut appliquer certains principes de base et de la logique aux besoins particuliers de chaque enquête.

Bien que trente principes individuels de l'élaboration d'un questionnaire soient présentés ici, ils ne doivent pas être considérés comme indépendants les uns des autres ou du contexte dans lequel une enquête se déroule. Le questionnaire est une partie intégrante de toute enquête — on ne saurait trop insister là-dessus. Puisque le questionnaire *ne peut pas* être isolé des divers autres aspects d'une enquête, le lecteur doit, *pendant* l'élaboration du questionnaire, examiner les rapports entre cet instrument et les objectifs de l'enquête, la population visée, la collecte, le codage, la saisie, la vérification et l'imputation des données, le caractère confidentiel des informations fournies et les essais préliminaires.

Comme cet exposé n'a pas été conçu comme un traité complet sur l'élaboration des enquêtes ou des questionnaires, certains lecteurs, selon leur point de vue sur les différents aspects d'une enquête, remarqueront peut-être des lacunes dans les principes proposés ou préféreront exclure certains principes parce qu'ils semblent relever davantage d'une autre étape que la conception du questionnaire.

Les principes de base de l'élaboration d'un questionnaire énoncés portent sur le contenu, la formulation, la présentation et les essais préliminaires. Le questionnaire influe beaucoup sur la possibilité pour une enquête d'atteindre ses objectifs. Contrairement aux autres grandes étapes d'une enquête, par exemple la construction du plan de sondage ou les opérations de traitement des données, le questionnaire touche directement le répondant. Il est donc essentiel que le contenu, la formulation et la présentation du questionnaire permettent d'obtenir du répondant des renseignements fiables, valables et correspondant aux informations recherchées. L'auteur reconnaît que bien que certains des principes semblent évidents quand ils

¹ Larry Swain, anciennement de la Division des méthodes de recensement et d'enquête-ménages, Statistique Canada, actuellement à la Division de la planification des ressources humaines, Commission de la Fonction publique, Ottawa, Canada KIA 0M7.

- SIEDULE, T., SKOULAS, N., et NEWTON, K. (1976). *La population active et les politiques économiques - une analyse économétrique*, Conseil économique du Canada, Ottawa.
- STATISTIQUE CANADA (1976). *Méthodologie de l'enquête sur la population active du Canada*. No de Catalogue 71-716.
- TIAO, G.C., et BOX, G.E.P. (1981). Modeling multiple time series with applications. *Journal of American Statistical Association*, 76, 802-816.
- TIAO, G.C., et TSAY, R.S. (1983). Multiple time series modeling and extended sample cross correlations. *Journal of Business and Economic Statistics*, 1, 43-56.
- ZELNER, A. (1979). Causality and econometrics. *Carnegie-Rochester Conference Series on Public Policy*, no. 10 (Karl Brunner et Allan H. Meltzer eds.), supp. du Journal of Monetary Economics, Amsterdam: North-Holland Publishing Company.

deux modèles de séries chronologiques multidimensionnelles suivant les méthodes élaborées par Tiao et Box (1981) et Tiao et Tsay (1983). Les résultats des modèles vectoriels ARMA concordent avec les résultats préliminaires des modèles unidimensionnels ARMMI (corrélations avec décalage par paire des résidus).

Le premier modèle vectoriel ARMA montre que la série BAC devance d'un mois la série NTC. De plus, le modèle indique une relation directe entre les termes de la série BAC et une relation inverse entre ceux de la série NTC.

Le second modèle vectoriel ARMA indique que la série JLO devance de deux mois la série BAC et qu'il y a une rétroaction d'un mois de BAC sur JLO. On observe, en outre, une relation directe entre les termes de la série BAC mais une relation inverse entre ceux de la série JLO. Le second modèle ARMA indique aussi que la série BAC devance la série JLO de deux mois.

Ces résultats empiriques, fondés sur des données couvrant la période 1975-1982, ne sont pas contraires à la théorie économique. Ils sont, en outre, conformes aux résultats d'études antérieures réalisées au Canada et fondées sur des données portant sur une période antérieure à 1975. Les résultats de ces études confirmaient l'affirmation générale selon laquelle l'assouplissement du programme d'assurance-chômage avait eu pour effet d'inciter un plus grand nombre de personnes, surtout des jeunes et des femmes adultes, à quitter leur emploi et avait contribué à accroître le niveau de chômage. Il semble donc que les révisions apportées au programme d'assurance-chômage en 1977 n'aient rien changé à la situation.

Il aurait été très intéressant d'analyser les effets de la forte récession du début des années 1980 sur les séries étudiées mais, compte tenu de la longueur de ces séries, la suppression de cette période de récession aurait réduit les séries à un point tel qu'elles n'auraient convenu à aucun modèle statistique.

BIBLIOGRAPHIE

BOX, G.E.P., et JENKINS, G.M. (1970). *Time Series Analysis Forecasting and Control*. San Francisco: Holden Day.

GRANGER, C.W.J. (1969): Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 424-438.

GREEN, C., et COUSINEAU, J.M. (1976). *Chômage et programmes d'assurance-chômage*. Conseil économique du Canada, Ottawa.

GRUBEL, M.G., MAKI, D., et SAX, S. (1975). Real and insurance induced unemployment in Canada. *Revue canadienne d'économique*, VIII, 174-191.

HAUGH, L.D. (1976). Checking the independence of two covariance stationary time series: A univariate residual cross-correlation approach. *Journal of American Statistical Association*, 71, 378-385.

JUMP, G.V., et REA, S.A. (1975). *The Impact of the 1971 Unemployment Insurance Act on Work Incentives and the Aggregate Labour Market*. Institute for Policy Analysis, Université de Toronto.

LAZAR, F. (1978). The impact of the 1971 unemployment insurance revisions on unemployment rates: Another look. *Revue canadienne d'économique*, août, 559-570.

LIU, L.M., HUDAK G.B. (1983). *Univariate - Multivariate Time Series and General Statistical Analysis*. Scientific Computing Associates, Illinois.

LJUNG, G.M., et BOX, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-307.

PIERCE, D.A., et HAUGH, L.D. (1977). Causality in temporal systems. *Journal of Econometrics*, 5, 265-293.

Les équations (4.10) et (4.11) indiquent l'existence d'une relation de rétroaction entre le nombre de personnes ayant perdu leur emploi et le nombre de bénéficiaires de l'assurance-chômage, cette relation étant comparable à celle qui a été mise à jour dans la section 3. La série JLo devance de deux mois la série BAC (équation 4.10) et le retard d'un mois qu'elle accuse par rapport à la série BAC (équation 4.11) influe sensiblement sur sa valeur courante. En outre, les valeurs des deux variables endogènes BAC et JLo sont influencées par leurs valeurs du mois précédent; positivement pour la série BAC et négativement pour la série JLo. Les coefficients de a'_{-6} et de a'_{-12} dénotent la relation que crée le mouvement saisonnier entre les deux séries. Le fait que la série JLe soit caractérisée par deux sommets accentués au cours d'une même année (un en hiver et un en été — voir figure 6) nous oblige à introduire un coefficient de moyenne mobile au décalage 6.

Ces observations ne viennent pas contredire la théorie économique. On a prétendu à bon droit que la causalité ne pouvait être uniquement déterminée par des résultats empiriques et qu'elle devait être aussi appuyée par la théorie économique (voir, par exemple Zellner 1979). Nous pouvons facilement admettre que, dans le contexte d'une récession économique, une augmentation du nombre de travailleurs licenciés se traduira par un nombre accru de bénéficiaires de l'assurance-chômage. En retour, un accroissement du nombre de ces bénéficiaires engendrera d'autres licenciements puisque, dans un tel contexte, la plupart des entreprises procèdent à des mises à pied temporaires pour pouvoir ensuite rappeler leurs employés lorsqu'une reprise s'amorce.

L'équation (4.12) soulève une question intéressante en indiquant que la série des bénéficiaires de l'assurance-chômage devance de deux mois la série des personnes ayant quitté leur emploi. Nous ne voyons pas à prime abord pourquoi il devrait en être ainsi. Nous pourrions trouver des explications plausibles de ce phénomène en analysant la dynamique à court terme des marchés du travail au Canada, mais une analyse approfondie de la question exigerait des données longitudinales. Nous pouvons toutefois avancer l'hypothèse, parmi tant d'autres, selon laquelle une augmentation du nombre de personnes ayant perdu leur emploi et, partant, du nombre de bénéficiaires de l'assurance-chômage incite d'autres membres de la famille à chercher un emploi pour combler la perte de revenu. Ces chercheurs d'emploi sont les nouveaux venus et les personnes qui reviennent sur le marché du travail après une certaine absence. Ces deux catégories de personnes peuvent très difficilement se trouver un emploi en période de récession, lorsque le nombre de licenciements est à la hausse. Ces groupes sont surtout composés de jeunes et de femmes de plus de 25 ans, qui sont prêts dans un premier temps à accepter n'importe quel emploi pourvu que cet emploi procure un revenu additionnel à la famille. Ces personnes pourraient travailler le temps qu'il faut pour qu'elles deviennent admissibles aux prestations d'assurance-chômage. Une fois admissibles à ces prestations, elles quitteraient leur emploi afin d'en chercher un autre plus conforme à leurs goûts et à leurs aptitudes.

5. CONCLUSIONS

Le principal objet de cette étude a été de déterminer s'il existait des relations temporelles entre la série des bénéficiaires de l'assurance-chômage (BAC) et les séries du niveau total de chômage (NTC), du nombre de personnes ayant perdu leur emploi (JLo) et du nombre de personnes ayant quitté leur emploi (JLe), en construisant des modèles dynamiques de séries chronologiques multidimensionnelles.

Nous avons commencé par faire une analyse préliminaire en tentant de découvrir des relations temporelles par paire entre les séries NTC, BAC, JLo et JLe, selon le modèle défini par Granger (1969) et Pierce et Haugh (1977). Les résultats de notre analyse ont révélé l'existence de relations entre les quatre variables étudiées. Nous avons ensuite défini et estimé

Estimations de paramètres pour les séries BAC, JLo et JLe transformées

$\bar{\theta}_1$	$\bar{\theta}_2$	$\bar{\theta}_3$	$\bar{\theta}_4$	$\bar{\theta}_5$	$\bar{\theta}_6$
$\begin{bmatrix} 0.617 & - & - \\ (0.086) & & \\ 0.577 & -0.285 & - \\ (0.099) & (0.096) & \\ -0.411 & & \\ (0.088) & & \end{bmatrix}$	$\begin{bmatrix} 0.303 & - & - \\ (0.083) & & \\ -0.403 & - & - \\ (0.084) & & \\ 0.268 & -0.080 & - \\ (0.080) & & \end{bmatrix}$	$\begin{bmatrix} 0.014 & 0.117 & 0.339 \\ (0.077) & & \\ 0.797 & - & \\ (0.096) & 0.525 & - \\ (0.094) & & \\ 0.831 & - & \\ (0.094) & & \end{bmatrix}$	$\begin{bmatrix} 0.014 & 0.117 & 0.339 \\ (0.077) & & \\ 0.797 & - & \\ (0.096) & 0.525 & - \\ (0.094) & & \\ 0.831 & - & \\ (0.094) & & \end{bmatrix}$	$\begin{bmatrix} 0.014 & 0.117 & 0.339 \\ (0.077) & & \\ 0.797 & - & \\ (0.096) & 0.525 & - \\ (0.094) & & \\ 0.831 & - & \\ (0.094) & & \end{bmatrix}$	$\begin{bmatrix} 0.014 & 0.117 & 0.339 \\ (0.077) & & \\ 0.797 & - & \\ (0.096) & 0.525 & - \\ (0.094) & & \\ 0.831 & - & \\ (0.094) & & \end{bmatrix}$

Matrices des corrélations de décalage +, -, et .

...
...
...
DÉCALAGES 1 À 6					
...
...
...
DÉCALAGES 7 À 12					
.. +
...
...
DÉCALAGES 13 À 18					
.. -
...
...
DÉCALAGES 19 À 24					
...
...
...

Pour éviter qu'il y ait multicollinéarité entre les séries NTC et JLo, nous avons défini deux modèles vectoriels ARMA: un modèle (1,2)(0,1)₁₂, qui met en relation le nombre de bénéficiaires de l'assurance-chômage et le niveau total de chômage, et un modèle (2,6)(0,1)₁₂, qui met en relation les bénéficiaires de l'assurance-chômage, les personnes ayant perdu leur emploi et celles ayant quitté leur emploi. Ces modèles ont été définis et estimés par la méthode du maximum de vraisemblance du programme de Scientific Computing Associates (Liu et Hudak 1983). Les deux modèles sont ajustés respectivement aux données originales transformées:

$$(4.6) \quad \begin{pmatrix} bac_t \\ ntc_t \end{pmatrix} = (1 - B)(1 - B_{12}) \log_{10} \begin{pmatrix} BAC_t \\ NTC_t \end{pmatrix}$$

et

$$(4.7) \quad \begin{pmatrix} bac_t \\ jlo_t \end{pmatrix} = (1 - B)(1 - B_{12}) \log_{10} \begin{pmatrix} BAC_t \\ JLo_t \end{pmatrix}$$

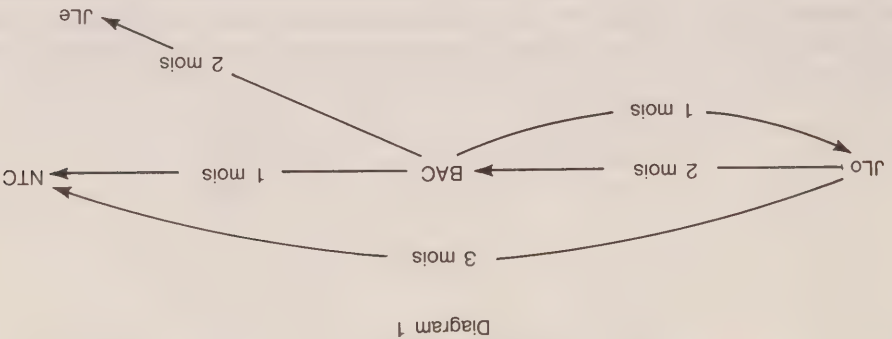
Le tableau 3 donne la valeur des paramètres du modèle vectoriel (1,2)(0,1) ainsi que l'ermatrice des variances-covariances des résidus, qui sont présentées au tableau 3 ne peuvent être comparées à celles des modèles unidimensionnels (tableau 1) parce que, dans le premier cas, le modèle a été ajusté aux données transformées normalisées alors que, dans le deuxième cas, le modèle a été ajusté aux données des résidus, nous permet de croire que le modèle est approprié. On désigne par un signe positif (+) une estimation qui est supérieure au double de son erreur type et par un signe négatif (-) une estimation qui est inférieure au double de son erreur type et on désigne par un point une valeur non significative fondée sur le critère précédent.

Tableau 3
Estimation de paramètres pour les séries BAC et NTC transformées

$\hat{\phi}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\bar{\Sigma}$
$\begin{bmatrix} 0.669 & -0.347 \\ (0.089) & (0.115) \end{bmatrix}$	$\begin{bmatrix} - & - \\ - & - \end{bmatrix}$	$\begin{bmatrix} - & - \\ (0.116) & 0.308 \end{bmatrix}$	$\begin{bmatrix} 0.429249 & 0.131532 \\ - & 0.544389 \end{bmatrix}$
$\bar{\theta}_1$	$\bar{\theta}_2$	$\bar{\theta}_1$	$\bar{\Sigma}$
$\begin{bmatrix} 0.794 & - \\ (0.090) & - \end{bmatrix}$	$\begin{bmatrix} - & - \\ - & - \end{bmatrix}$	$\begin{bmatrix} - & - \\ (0.086) & 0.705 \end{bmatrix}$	
$\bar{\theta}_{12}$			

3) On observe une relation unidirectionnelle entre les séries BAC et JLe de sorte que la première devancerait la seconde de 2 mois. On observe toutefois au décalage 6 l'effet d'une rétroaction tardive qui s'explique par le fait que la série JLe est caractérisée par deux sommets dans la même année, un en hiver et un en été, comme l'indique la figure 6. Enfin, on observe une forte relation unidirectionnelle et instantanée entre les séries JLo et NTC de sorte que la première devancerait la seconde de 3 mois.

Ces observations nous ont permis de tracer le diagramme 1 qui servira à définir un modèle, plus complexe, de séries chronologiques multidimensionnelles, qui tient compte des relations partielles entre les variables.



4. CONSTRUCTION D'UN MODÈLE DE SÉRIES CHRONOLOGIQUES MULTI-DIMENSIONNELLES POUR LES BÉNÉFICIAIRES DE L'ASSURANCE-CHÔMAGE, LE NIVEAU TOTAL DE CHÔMAGE, LES PERSONNES AYANT PERDU LEUR EMPLOI ET LES PERSONNES AYANT QUITTÉ LEUR EMPLOI

Dans la section précédente, nous avons conclu qu'il existait des relations par paire, au sens où l'entendent Granger (1969) et Pierce et Haugh (1977), entre les quatre variables étudiées. Compte tenu de ces observations, nous voulons, dans la présente section, définir et estimer deux modèles de séries chronologiques multidimensionnelles à l'aide des méthodes élaborées par Tiao et Box (1981) et Tiao et Tsay (1983). Ces modèles serviront à expliquer la dynamique des variables en question.

Nous définissons comme suit un modèle vectoriel ARMA pour les séries saisonnières:

$$(4.1) \quad \bar{\phi}(B)\bar{\Phi}(B^s)\bar{Z}_i' = \bar{\theta}(B)\bar{\Theta}(B^s)\bar{a}_i'$$

où

$$(4.2) \quad \bar{\phi}(B) = \bar{I} - \bar{\phi}_1 B - \dots - \bar{\phi}_p B^p$$

$$(4.3) \quad \bar{\Phi}(B^s) = \bar{I} - \bar{\Phi}_1 B^s - \dots - \bar{\Phi}_p B^{sp}$$

$$(4.4) \quad \bar{\theta}(B) = \bar{I} - \bar{\theta}_1 B - \dots - \bar{\theta}_q B^q$$

$$(4.5) \quad \bar{\Theta}(B^s) = \bar{I} - \bar{\Theta}_1 B^s - \dots - \bar{\Theta}_q B^{sq}$$

sont les polynômes de matrice en B (l'opérateur de décalage défini par $B^m Z_i' = Z_{i-m}'$), les $\bar{\phi}$'s, $\bar{\Phi}$'s, $\bar{\theta}$'s et $\bar{\Theta}$'s sont des matrices $k \times k$, s est le mouvement périodique saisonnier, a_i' est une suite de vecteurs aléatoires de changement $IID N(\bar{0}, \bar{\Sigma})$ et \bar{Z}_i' est un vecteur de séries chronologiques stationnaires.

Tableau 1

Modèles ARMMI unidimensionnels			Modèles ARMMI		
Séries	Q(24)		Q(24)		σ^2
Bénéficiaires de l'assurance-chômage (BAC)	$(1 - 0.68B) \Delta \Delta_{12} \log_{10} BAC_t = (1 - 0.80B_{12}) a_t$		11.55		0.000140
Niveau total	$(1 - 0.25B_3) \Delta \Delta_{12} \log_{10} NTC_t = (1 - 0.84B_{12}) a_t$		9.13		0.000395
de chômage (NTC)	$(1 - 0.31B_3) \Delta \Delta_{12} \log_{10} JLo_t = (1 - 0.67B_{12}) a_t$		15.78		0.000604
Personnes ayant perdu leur emploi (JLo)	$(1 - 0.37B_3) \Delta \Delta_{12} \log_{10} JLe_t = (1 - 0.40B - 0.25B_2) (1 - 0.87B_{12}) a_t$		14.58		0.000627
Personnes ayant quitté leur emploi (JLe)					

Tableau 2

Corrélation avec décalage entre la série des bénéficiaires de l'assurance-chômage, et la série du niveau total de chômage et ses deux grandes composantes, personnes ayant perdu leur emploi et personnes ayant quitté leur emploi

DÉCALAGES	k	$BAC_{(t-k)} - NTC_t$	$BAC_{(t-k)} - JLo_t$	$BAC_{(t-k)} - JLe_t$	$JLo_{(t-k)} - NTC_t$
-6	-0.07	0.01	0.27 ^a	0.15	
-5	0.05	-0.04	-0.09	-0.04	
-4	-0.09	0.03	-0.08	-0.01	
-3	0.01	0.01	-0.11	-0.06	
-2	0.14	0.28 ^a	0.14	-0.01	
-1	0.14	0.08	-0.01	0.21	
0	0.16	0.32 ^b	0.06	0.39 ^b	
1	0.22 ^a	0.29 ^a	0.04	0.14	
2	0.12	0.12	0.26 ^a	-0.16	
3	-0.07	0.00	0.19	0.42 ^b	
4	0.12	-0.05	0.01	-0.05	
5	-0.06	0.09	0.11	-0.04	
6	0.13	0.00	0.08	0.05	

^a seuil de signification de 10%.

^b seuil de signification de 50%.

^a seuil de signification de 5%.

^b seuil de signification de 1%.

Dans la présente section, nous nous livrons à une analyse préliminaire en tentant de

de découvrir des relations temporelles par paire entre le niveau total de chômage, le nombre de bénéficiaires de l'assurance-chômage, le nombre de personnes ayant perdu leur emploi et le nombre de celles ayant quitté leur emploi. L'existence de telles relations permettra de construire un modèle de séries chronologiques multidimensionnelles, qui servira à expliquer la dynamique des variables énumérées ci-dessus.

Les relations par paire entre les séries NTC, BAC, JLo et JLe sont déterminées au moyen des corrélations avec retard des résidus ou *innovations* des modèles ARMMI (Box et Jenkins, 1970) qui donnent une bonne approximation des données. Plusieurs auteurs dont Pierce et Haugh (1977) ont fait valoir avec raison que les corrélations avec décalage entre les bruits blancs obtenus au moyen de filtres différents sont telles qu'elles favorisent l'acceptation de l'hypothèse nulle de l'indépendance, alors que celle-ci n'existe pas. Pierce et Haugh (1977) proposent d'utiliser des modèles de régression dynamiques. Il faut toutefois, pour cela, déterminer quelle variable est la *cause* et quelle variable est l'*effet*. Pour l'instant, nous cherchons seulement à savoir s'il existe une relation temporelle dans chaque paire de variables analysée. Le tableau 1 donne les modèles ARMMI ajustés à chaque série, leurs paramètres estimés au moyen des moindres carrés inconditionnels, les résultats du test du *portmanteau* (Ljung et Box, 1978) et la variance résiduelle.

Les valeurs du paramètre statistique Q se trouvent dans l'intervalle d'acceptation de l'hypothèse nulle, selon laquelle les résidus présentent un caractère aléatoire dans chaque cas. Comme, toutefois, ce test est appliqué à un ensemble d'autocorrélations de résidus pour divers décalages, il pourrait ne pas révéler l'existence d'une forte autocorrélation pour un certain décalage k . Par conséquent, nous avons aussi vérifié s'il existait une autocorrélation de résidus pour chaque décalage. Pour tester la variance de l'autocorrélation, nous sommes servis d'une formule d'approximation plus précise que $1/N$ pour les petits échantillons, c'est-à-dire $(N - |k|)N^{-2}$ tel que la définit Haugh (1976).

Ayant obtenu des résultats satisfaisants avec ces modèles, nous avons calculé le coefficient de corrélation avec décalage $f_{xy}(k)$ entre les séries analysées. On se sert du paramètre statistique $S_M^*(k)$ (Haugh 1976) pour tester l'indépendance des séries en supposant que les résidus sont distribués suivant une loi normale que $E[f_{xy}(k)] = 0$ et $\text{Var}[f_{xy}(k)] = (N - |k|)N^{-2}$ et que

$$S_M^* = N^2 \sum_{k=-M}^M (N - |k|)^{-1} f_{xy}(k)^2$$

suit une distribution de *khi-carre* avec $2M + 1$ degrés de liberté. Afin de déterminer la direction des relations par paire, nous avons modifié le paramètre statistique S_M^* , qui n'a été calculé que pour des valeurs positives ou négatives de k , c'est-à-dire à l'exclusion de la valeur nulle. Le tableau 2 donne les estimations de la corrélation avec décalage entre la série des bénéficiaires de l'assurance-chômage (BAC) et la série du niveau total de chômage (NTC) et ses deux principales composantes, les personnes ayant perdu leur emploi (JLo) et les personnes ayant quitté leur emploi (JLe). Certaines valeurs du tableau sont marquées d'un (a) ou d'un (b) suivant qu'elles sont significatives à un seuil de confiance de 5% ou de 1% respectivement. En ce qui concerne les séries BAC et JLo, nous avons calculé S_M^* pour les valeurs de k allant de ± 1 à ± 6 et de moins ± 1 à ± 2 pour déterminer s'il existait une relation unidirectionnelle dominante. Les résultats ont indiqué qu'il n'y avait aucune direction dominante entre chaque paire de variables mais qu'il existait une relation de rétroaction entre celles-ci.

Les résultats du tableau 2 peuvent être résumés comme suit :

- 1) Les données indiquent une relation unidirectionnelle entre les séries BAC et NTC de sorte que BAC devancerait NTC d'un mois.
- 2) Il y a rétroaction entre les séries BAC et JLo et cette rétroaction est caractérisée par une forte relation instantanée. Compte tenu du décalage entre les deux variables, la rétroaction semble être amorcée par la série JLo au décalage 2.

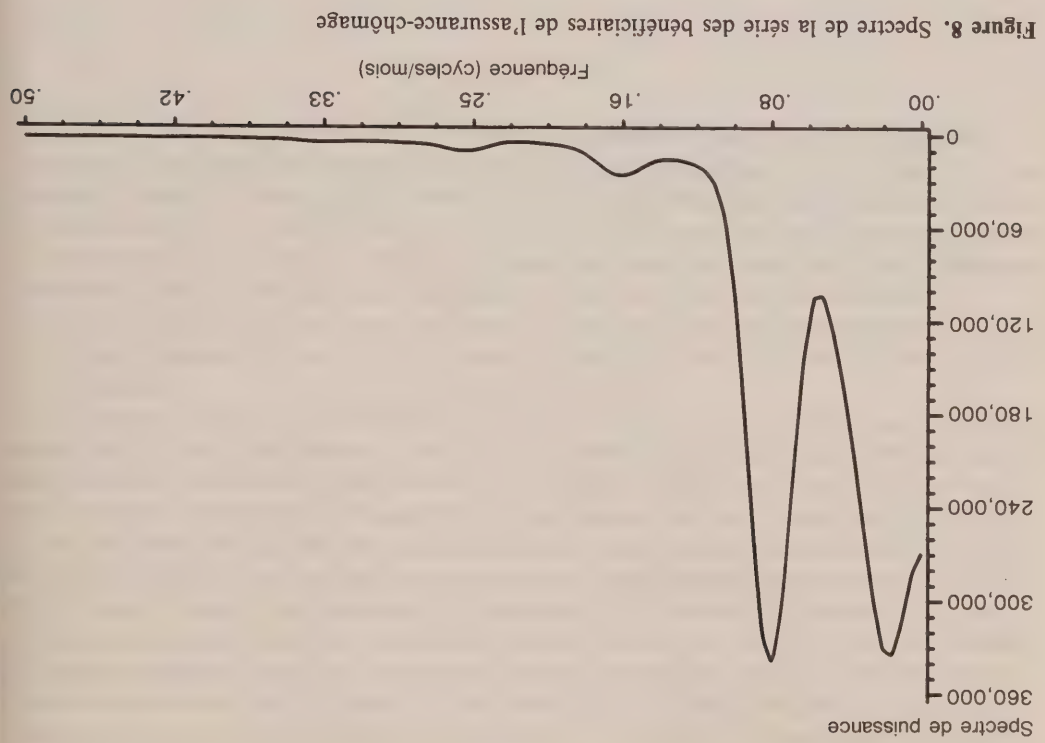


Figure 8. Spectre de la série des bénéficiaires de l'assurance-chômage

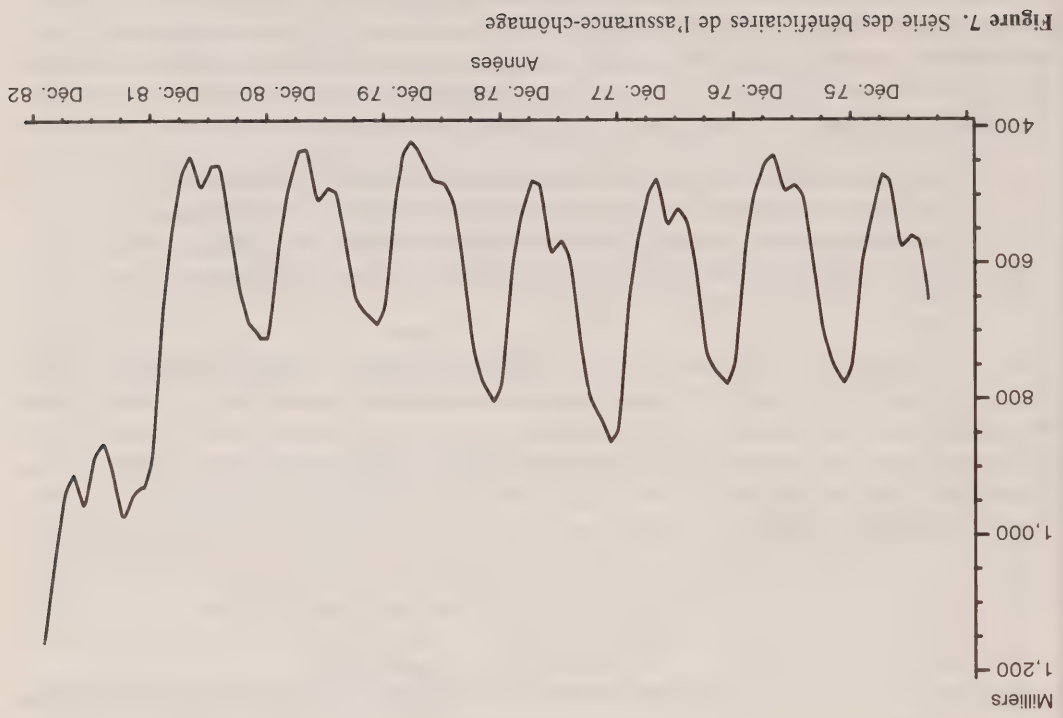
La figure 8 donne le spectre de la série des bénéficiaires de l'assurance-chômage. On y observe une très forte puissance à la fréquence 0.0167, qui correspond à un cycle de 60 mois, ainsi qu'aux fréquences associées à la bande saisonnière fondamentale. Les variations saisonnières comptent pour beaucoup plus dans la variance totale de la série qu'elles ne le faisaient pour la série NTC et ses deux grandes composantes. Enfin, les fluctuations irrégulières sont négligeables par rapport à la tendance-cycle et aux composantes saisonnières.

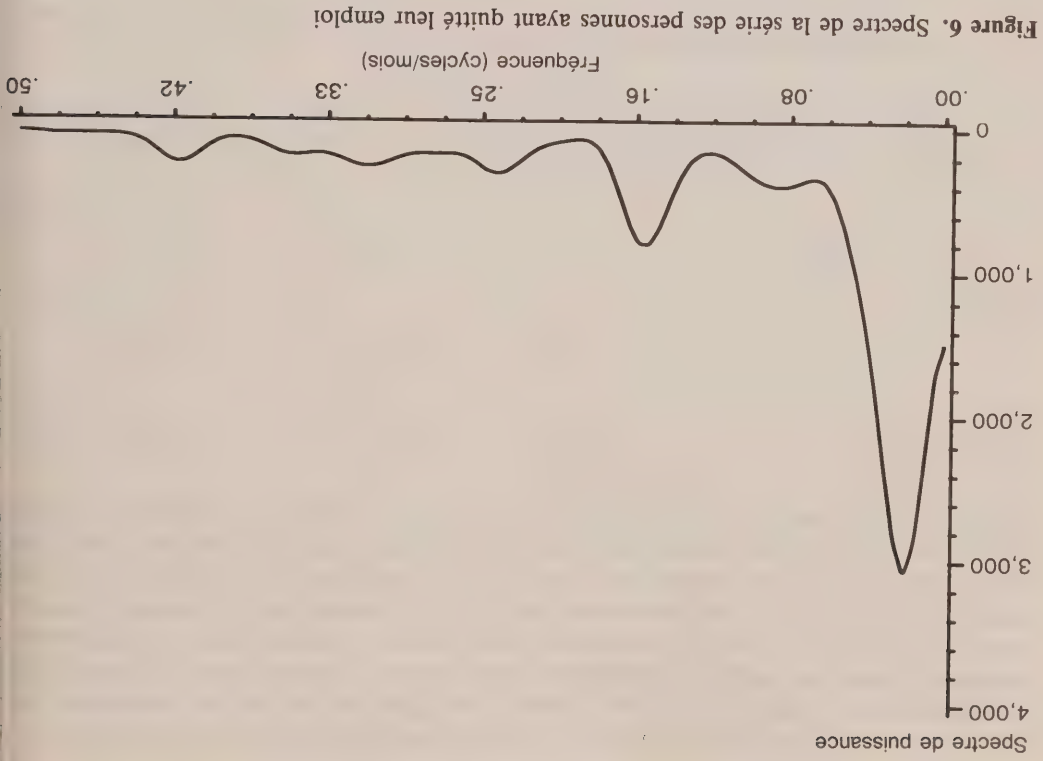
3. RELATIONS PAR PAIRE ENTRE LE NOMBRE DE BÉNÉFICIAIRES DE L'ASSURANCE-CHÔMAGE, LE NIVEAU TOTAL DE CHÔMAGE, LE NOMBRE DE PERSONNES AYANT PERDU LEUR EMPLOI ET LE NOMBRE DE PERSONNES AYANT QUITTÉ LEUR EMPLOI

Plusieurs études réalisées au Canada il y a déjà un certain nombre d'années (par exemple, Grubel et coll. 1975; Green et Cousineau 1976; Jump et Rea 1975; et Siedule et coll. 1976) viennent confirmer l'affirmation générale selon laquelle le niveau de chômage. Lazar (1978) d'assurance-chômage en 1971 a eu tendance à accroître le niveau de chômage. Lazar (1978) montre que les modifications de 1971 ont eu pour effet d'accroître la durée du chômage et d'inciter un nombre proportionnellement plus élevé de personnes, surtout chez les jeunes et les femmes adultes, à quitter leur emploi. Ces études ont été faites avant les modifications de 1975 qui visaient à révaloriser les formes d'encouragement au travail. On espérait que les modifications apportées après 1975 supprimaient les effets nuisibles du programme sur le niveau total de chômage.

bénéficiaires de l'assurance-chômage représente une proportion relativement stable et importante du nombre total de chômeurs établi selon l'EPA. Il convient toutefois de souligner qu'à cause d'une différence de définitions, certaines personnes sont considérées comme des chômeurs dans l'EPA mais ne figurent pas dans les dossiers de l'A-C. Il s'agit, notamment, des nouveaux venus et des personnes qui reviennent sur le marché du travail, de toutes les personnes qui n'ont pas travaillé suffisamment longtemps pour avoir droit à des prestations et des personnes en chômage qui étaient auparavant des travailleurs autonomes. En revanche, certaines catégories de personnes qui sont admissibles aux prestations d'assurance-chômage ne sont pas considérées comme des chômeurs selon l'EPA. Mentionnons, par exemple, les pêcheurs indépendants, au cours de la saison morte, les femmes qui bénéficient d'un congé de maternité et les employés qui sont absents de leur poste pour cause de maladie ou d'invalidité.

La série des bénéficiaires de l'A-C (sans rémunération) est un indicateur sensible de la conjoncture du marché du travail. Elle trace un portrait fidèle de la main-d'œuvre qui a récemment quitté le marché du travail et qui bénéficie de l'assurance-chômage. Comme on peut le voir à la figure 7, la série originale des bénéficiaires de l'assurance-chômage connaît de fortes fluctuations saisonnières; son mouvement est caractérisé par des sommets durant les mois d'hiver, lorsque les mauvaises conditions atmosphériques limitent le travail à l'extérieur dans des secteurs comme la pêche, la construction et la coupe de bois d'œuvre, ce qui provoque une forte hausse du nombre de demandes de prestations.





2.2 Bénéficiaires de l'assurance-chômage (BAC)

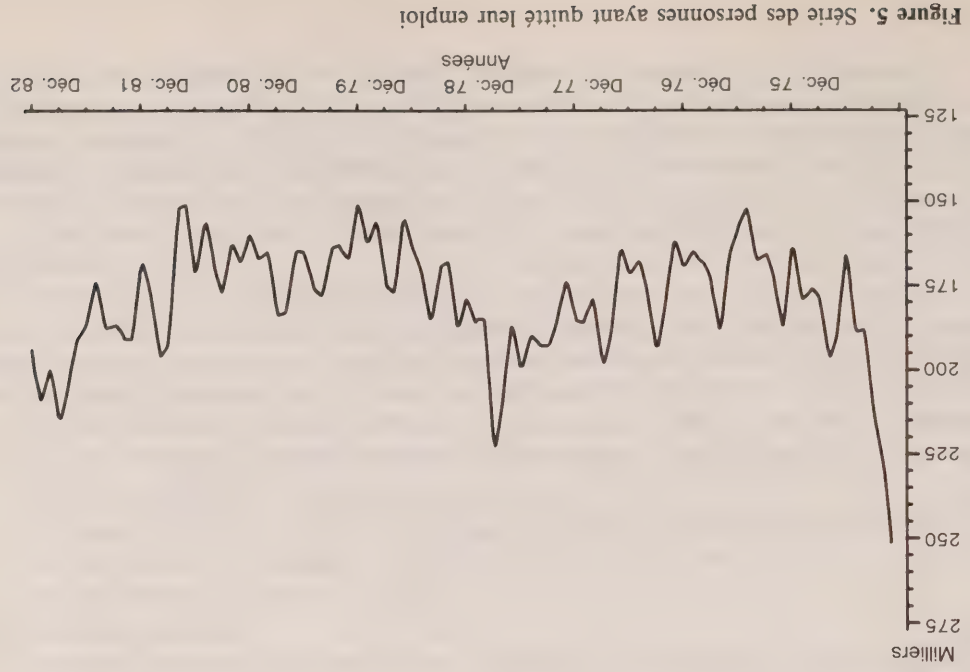
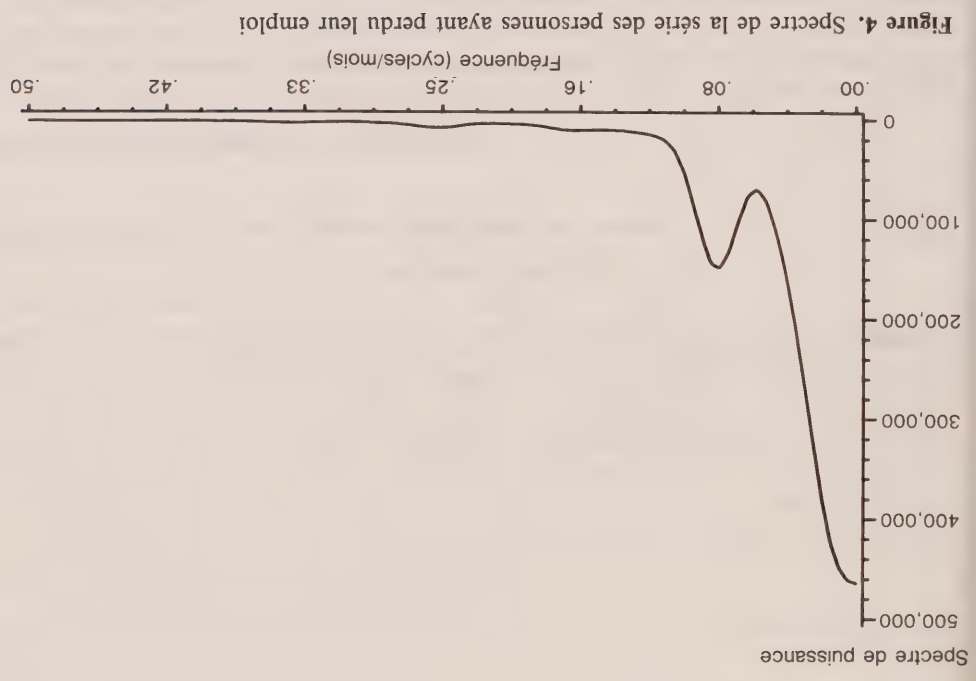
Les données mensuelles concernant les bénéficiaires de l'assurance-chômage couvrent toutes les personnes qui reçoivent des prestations pour une semaine déterminée, notamment la semaine de référence de l'EPA. Il ne s'agit pas d'un échantillon puisque les données concernent l'ensemble des bénéficiaires. L'A-C vise pratiquement tous les membres de la population active qui reçoivent une rémunération ainsi que les membres des forces armées. Les principales exceptions sont :

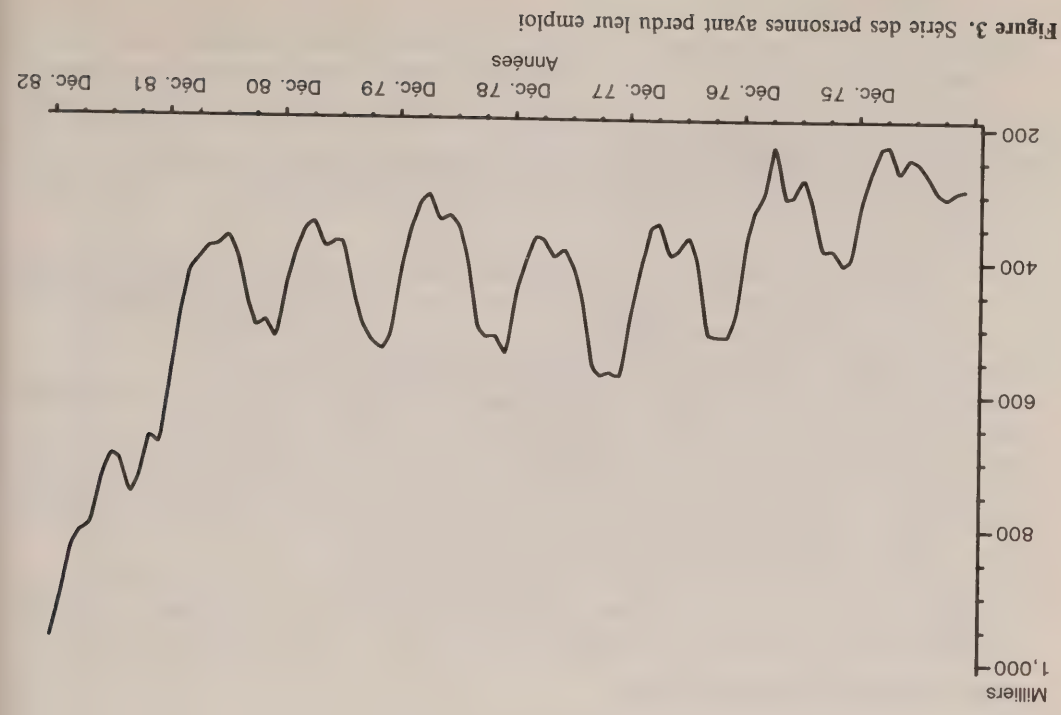
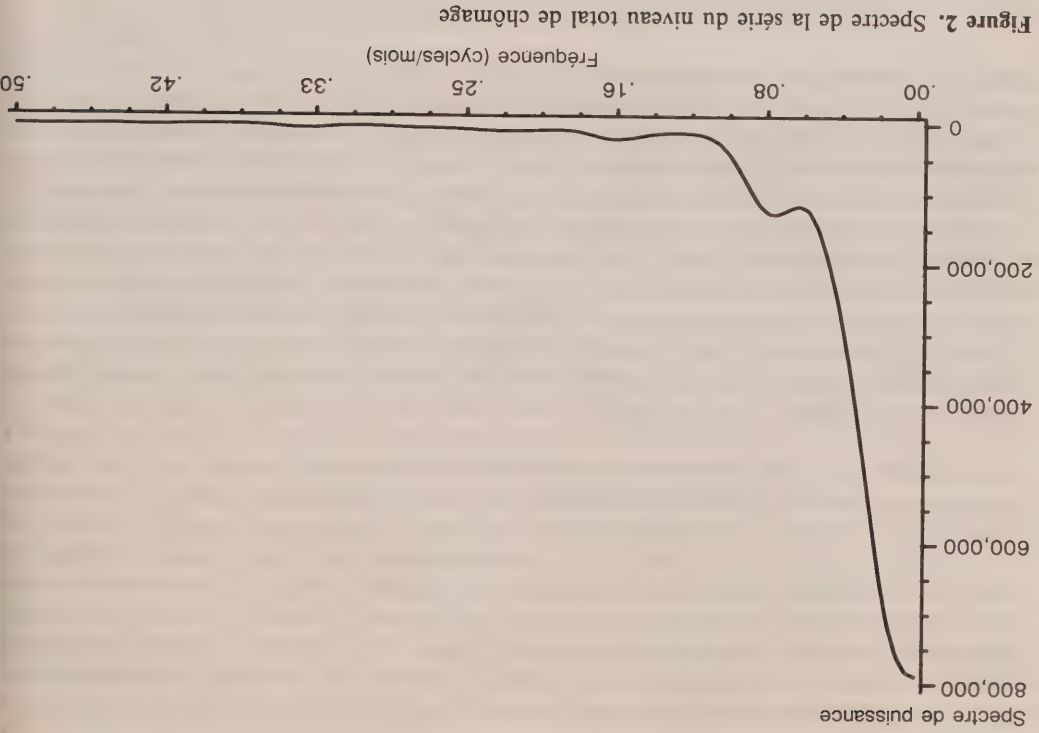
- les personnes de 65 ans et plus;
- Les personnes qui travaillent moins de 15 heures par semaine;
- Les personnes qui gagnent moins de 20% de la rémunération hebdomadaire assurable maximum (en 1982, ce maximum était de \$70).

Pour être admissible à des prestations, une personne doit être disponible pour travailler et être apte à occuper un emploi, capable de se trouver un emploi convenable et satisfaisant aux conditions d'admissibilité. Auparavant, huit semaines de travail suffisaient à une personne pour qu'elle soit admissible à des prestations mais, depuis décembre 1977, ce nombre varie de 10 à 14 semaines selon le taux de chômage enregistré dans la région où le prestataire a son domicile. Les prestations commencent à être versées après une période de carence de deux semaines.

Les personnes admissibles à des prestations d'assurance-chômage peuvent gagner un revenu d'emploi n'excédant pas 25% du montant de leurs prestations et continuer de recevoir ces prestations. Cependant, ces personnes sont considérées comme occupées dans l'EPA. Si nous voulons déterminer la relation entre le nombre de bénéficiaires de l'assurance-chômage et le niveau total de chômage, il serait donc plus convenable d'utiliser la série des bénéficiaires de l'assurance-chômage qui ne reçoivent aucune rémunération. Ce sous-ensemble des

La figure 5 illustre la série des personnes ayant quitté leur emploi, laquelle est caractérisée par deux creux, un pour les mois d'hiver et l'autre pour les mois d'été. Le spectre de cette série est décrit à la Figure 6. Il indique que la puissance aux basses fréquences est surtout concentrée au niveau des fréquences correspondant au cycle économique, comme l'indique le sommet observé à la fréquence 0.022 (laquelle correspond à un cycle de 45 mois). En outre, les variations saisonnières sont concentrées autour de la première bande harmonique, ce qui confirme le fait que cette série présente deux creux saisonniers. Enfin, l'effet des fluctuations irrégulières sur la variance totale est plus notable que dans les deux séries précédentes.





La série du niveau total de chômage regroupe les personnes ayant perdu leur emploi (JLo), celles ayant quitté leur emploi (JLe), les nouveaux venus sur le marché du travail, les personnes qui reviennent sur le marché du travail après un an ou moins et celles qui y reviennent après plus d'un an (Statistique Canada 1976). Les deux premiers groupes sont ceux qui comptent le plus pour notre étude puisque leurs membres sont admissibles à des prestations et représentent environ 70% de tous les chômeurs.

Comme il n'est pas possible d'obtenir de données sur l'augmentation du nombre de chômeurs pour la période précédant l'année 1975, toutes les séries analysées couvrent la période de janvier 1975 à décembre 1982 et comprennent, par conséquent, les données les plus récentes possibles.

La figure 1 illustre la série originale du niveau total de chômage, laquelle est caractérisée par un sommet dans les mois d'hiver et par un creux dans les mois d'été. La figure 2 décrit le spectre de cette série. On y observe une puissance élevée entre les fréquences 0.00 et 0.05, cette dernière équivalant à un cycle économique ($f = 0.05$ correspond à un cycle de 20 mois). De même, la puissance est relativement élevée autour de la fréquence saisonnière fondamentale (0.083) mais est moindre aux bandes harmoniques. Enfin, les fluctuations irrégulières comptent pour peu dans la variance totale par rapport aux deux autres composantes. La figure 3 montre la série originale des personnes ayant perdu leur emploi et la Figure 4 décrit le spectre correspondant. Comme dans le cas de la série NTC, une forte puissance est observée aux fréquences correspondant à des cycles économiques mais, cette fois, la puissance saisonnière est concentrée dans la bande saisonnière fondamentale et est presque absente dans les bandes harmoniques. L'effet des fluctuations irrégulières est encore plus faible dans cette série.

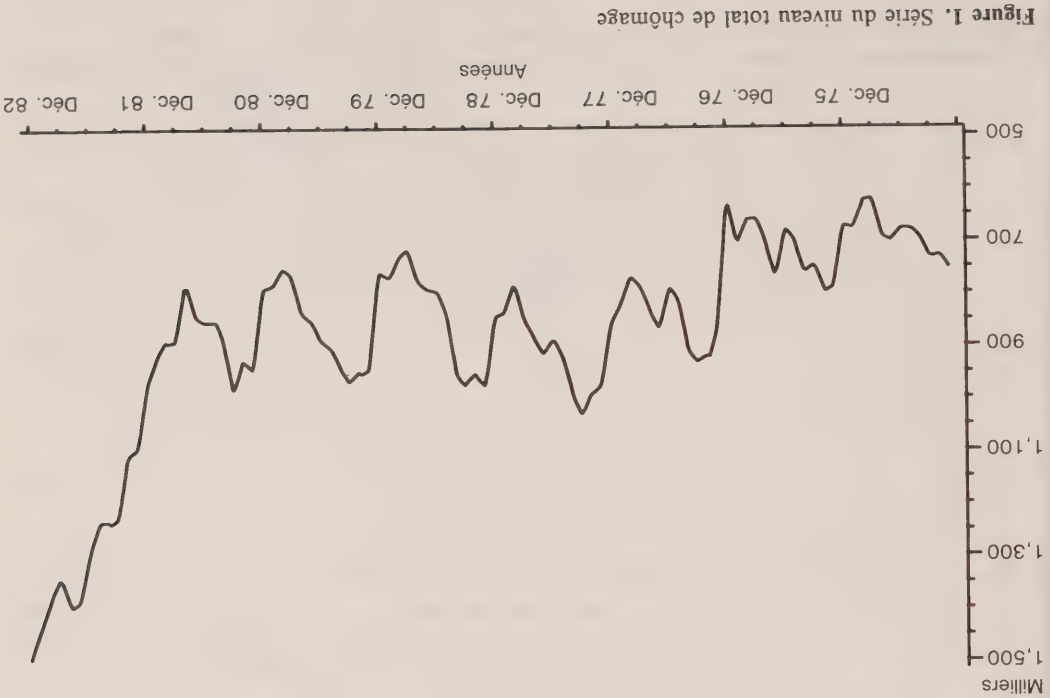


Figure 1. Série du niveau total de chômage

ayant perdu leur emploi représentent une plus forte proportion du nombre total de chômeurs. Comme la plupart des prestataires de l'assurance-chômage sont, de fait, des personnes qui ont perdu leur emploi, la proportion des bénéficiaires de l'assurance-chômage se trouve augmentée par rapport au niveau total de chômage.

La présente étude vise à vérifier s'il existe une relation temporelle entre le nombre de bénéficiaires de l'assurance-chômage (BAC) et le niveau total de chômage au Canada (NTC). L'analyse tient compte également des personnes ayant perdu leur emploi (JLo) et de celles ayant quitté leur emploi (JLe), deux groupes qui ont droit à des prestations et qui représentent la grande majorité des chômeurs au pays. L'existence d'une relation étroite entre ces variables pourrait servir à expliquer le comportement des marchés du travail. En outre, si une telle relation existe, elle peut nous amener à en découvrir d'autres du même genre, qui serviraient à estimer le niveau de chômage dans les petites régions où la taille de l'échantillon de l'enquête sur la population active est insuffisante. Dans la section 2, nous définissons chacune des quatre séries étudiées et analysons leurs principales caractéristiques à partir de leurs spectres respectifs. Dans la section 3, nous estimons, pour plusieurs décalages, les corrélations avec décalage des résidus des séries affectées d'un bruit blanc pour déterminer s'il existe des relations par paire et définir leur direction, s'il y a lieu. Les résidus sont calculés à l'aide de modèles ARMMI ajustés à chaque série. Dans la section 4, nous poursuivons les analyses des sections précédentes en définissant et en estimant deux modèles de séries chronologiques multidimensionnelles afin de comprendre la corrélation dynamique entre 1) les séries BAC et NTC et 2) les séries BAC, JLo et JLe. Enfin, la section 5 présente les principales conclusions de cette étude.

2. PRINCIPALES CARACTÉRISTIQUES DES SÉRIES ANALYSÉES

Pour comprendre le genre de relations qui existent entre la série BAC et la série NTC et les principales composantes, JLo et JLe, nous commençons par définir ces séries et analyser leurs principales caractéristiques à l'aide de leurs spectres.

2.1 Niveau total de chômage (NTC)

La Division de l'enquête sur la population active (EPA) de Statistique Canada recueille des données mensuelles au moyen d'une enquête menée auprès de 56,000 ménages représentatifs répartis dans tout le pays. Malgré les améliorations qui ont été apportées depuis 1952, l'EPA a subi d'importantes modifications à partir de 1976.

Les estimations du nombre de personnes occupées, du nombre de chômeurs et du nombre d'inactifs portent sur la semaine de référence de l'enquête mensuelle, habituellement la semaine incluant le 15^e jour du mois. L'échantillon de l'enquête représente, à quelques exceptions près, toutes les personnes de 15 ans et plus qui résident au Canada.

La population active comprend les personnes qui avaient un emploi ou étaient en chômage pendant la semaine de référence. Les personnes occupées sont celles qui :

- ont fait un travail quelconque;
- avaient un emploi mais n'étaient pas au travail pour l'une des raisons suivantes: maladie ou invalidité, mauvais temps, conflit de travail, vacances ou obligations personnelles ou familiales.

Le groupe des chômeurs comprend les personnes qui :

- étaient sans travail mais avaient activement cherché du travail au cours des quatre dernières semaines et étaient prêtes à travailler;
- n'avaient pas activement cherché de travail au cours des quatre dernières semaines mais avaient été mises à pied depuis 26 semaines ou moins et étaient prêtes à travailler;
- n'avaient pas activement cherché de travail au cours des quatre dernières semaines ou moins et étaient prêtes à travailler.

Formes de relation entre le niveau total de chômage et le nombre de bénéficiaires de l'assurance-chômage au Canada

ESTELA BEE DAGUM, GUY HUOT, NAZIRA GAIT,
et NORMAND LANIÉL¹

RÉSUMÉ

La présente étude vise à vérifier à l'aide de séries chronologiques unidimensionnelles et multidimensionnelles s'il existe des relations temporelles entre le nombre de bénéficiaires de l'assurance-chômage, le niveau total de chômage, le nombre de personnes ayant perdu leur emploi et le nombre de personnes ayant quitté leur emploi au Canada. Les résultats indiquent que, pour la période 1975-1982, la série des bénéficiaires de l'assurance-chômage devance d'un mois la série du niveau total de chômage et de deux mois celle des personnes ayant quitté leur emploi. Par ailleurs, on constate une relation de rétroaction entre la série des bénéficiaires de l'assurance-chômage et celle des personnes ayant perdu leur emploi.

MOTS CLÉS: Personnes ayant perdu leur emploi; personnes ayant quitté leur emploi; ARMMI; modèles vectoriels ARMA; séries chronologiques multidimensionnelles.

1. INTRODUCTION

L'assurance-chômage (A-C) joue un rôle fondamental en permettant aux marchés du travail du pays de s'adapter aux variations du niveau de production et du niveau d'emploi causées par les fluctuations de la demande et des échanges commerciaux. Dans le cadre plus global d'une politique du marché du travail, l'A-C a pour principal objet d'assurer une protection financière convenable aux personnes qui sont temporairement en chômage pour leur permettre de vivre plus facilement cette période de transition. Comme l'A-C élimine la principale menace découlant du chômage, c'est-à-dire la perte de revenus, les chercheurs d'emploi ne se sentent plus obligés de céder aux pressions économiques en acceptant des emplois qui ne conviennent pas à leurs compétences ou à leurs aptitudes. En permettant une recherche d'emploi plus systématique ou plus variée, l'A-C favorise une réaffectation efficiente des ressources humaines. En outre, lorsqu'il y a des fermetures temporaires d'usines, l'A-C atteint son objectif en accordant une protection financière aux travailleurs touchés, de telle sorte que l'employeur peut rappeler à tout moment la même main-d'œuvre expérimentée dont il bénéficiait auparavant. À la réouverture de l'usine, il n'a donc pas besoin de recruter et de former de nouveaux employés et évite, par le fait même, les dépenses inhérentes à ces activités. L'A-C permet en outre à l'employé de traverser sans trop d'ennuis la période de chômage et, partant, d'échapper aux difficultés financières.

Peu importe la situation, l'A-C doit être suffisamment flexible pour tenir compte de la conjoncture économique du moment, laquelle peut restreindre le nombre d'emplois disponibles et prolonger la période de chômage des chercheurs d'emploi. Dans le programme canadien d'assurance-chômage, cette flexibilité se traduit par une prolongation de la durée des prestations lorsque les taux de chômage dans les régions sont à la hausse.

L'écart entre la série du niveau total de chômage et celle des bénéficiaires de l'assurance-chômage tend à s'amenuiser en période de récession et à s'élargir en période de reprise. Lors- que la conjoncture économique se détériore et que les mises à pied surviennent, les personnes

¹ E. B. Dagum et G. Huot, Division des séries chronologiques recherche et analyse, Statistique Canada, N. Gait, Université de Sao Paulo, Brésil, cette dernière effectuait un séjour à Statistique Canada lorsque l'article a été écrit, et N. Lanier, Division des séries chronologiques recherche et analyse maintenant avec la Division des méthodes d'enquêtes-entreprises.

- WALLIS, K.F. (1974). Seasonal Adjustment and Relations Between Variables. *Journal of the American Statistical Association*, 69, 18-31.
- WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: Linear filters for the X-11 Method. *Journal of the Royal Statistical Society, série A*, 145, 74-85.
- YOUNG, A.H. (1968). Linear approximations to census and BLS seasonal adjustment methods. *Journal of the American Statistical Association*, 63, 445-457.

Compte tenu des observations précédentes, nous pouvons affirmer que la méthode de désaisonnalisation officielle de Statistique Canada produira les meilleures estimations possibles en période de récession.

REMERCIEMENT

Les auteurs tiennent à exprimer leur reconnaissance aux deux arbitres qui, par leurs précieux conseils, ont contribué à parfaire cette étude.

BIBLIOGRAPHIE

- BOX, G.E.P., et JENKINS, G.M. (1970). *Times Series Analysis: Forecasting and Control*. San Francisco: Holden Day.
- DAGUM, E.B. (1978). *Comparison and Assessment of Seasonal Adjustment Methods for Labor Force Series*. Washington, D.C.: U.S. Government Printing Office.
- DAGUM, E.B. (1980). *La méthode de désaisonnalisation X-11-ARMMI*. N° 12-564F au catalogue, Ottawa, Canada: Statistique Canada.
- DAGUM, E.B. (1982a). Revision of time varying seasonal filters. *Journal of Forecasting*, 1, 173-187.
- DAGUM, E.B. (1982b). The effects of asymmetric filters on seasonal factor revisions. *Journal of the American Statistical Association*, 77, 732-738.
- DAGUM, E.B., et MORRY, M. (1982). L'estimation des variations saisonnières dans les indices des prix à la consommation. Actes du colloque sur *La mesure du niveau des prix*. N° 22-24 au catalogue, Ottawa, Canada: Statistique Canada.
- HIGGINSON, J. (1977). Manuel d'utilisation du test de Bell Canada relatif à la sélection de schémas de composition des séries. Document de recherche n° 77-01-001, Groupe de la désaisonnalisation et des séries chronologiques, Statistique Canada.
- KENNY, P., et DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic time series. *Journal of the Royal Statistical Society, Série A*, 145, 1-41.
- KUIPER, J. (1978). A survey and comparative analysis of various methods of seasonal adjustment. *Seasonal Analysis of Economic Time Series*, (Ed. Arnold Zellner), Washington, D.C.: U.S. Government Printing Office, 59-76.
- KUIPER, J. (1981). The treatment of extreme values in the X-11-ARIMA program. *Time Series Analysis and Forecasting*, (Eds. Anderson, O. et Perrymann, M.R.), Amsterdam: North-Holland Publishing Co., 257-266.
- McKENZIE, S. (1982). An evaluation of concurrent adjustment on Census Bureau time series. *Proceedings of the Business and Economics Section of the American Statistical Association*.
- MORRY, M. (1975). A test for model selection. Document de recherche n° 75-12-016, Groupe de la désaisonnalisation et des séries chronologiques, Statistique Canada.
- OTTO, M. (1985). Effects of forecasts on the revisions of seasonally adjusted values using the X-11 seasonal adjustment procedure. *Proceedings of the Business and Economics Section of the American Statistical Association* (à paraître).
- PIERCE, D. (1980). Data revision with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.
- PIERCE, D., et McKENZIE S. (1985). On concurrent seasonal adjustment. Document technique, U.S. Bureau of the Census.
- SHISKIN, J., YOUNG, A.H., et MUSGRAVE, J.C. (1967). The X-11 variant of census method II seasonal adjustment program. Document technique n° 15, Washington, D.C.: U.S. Bureau of Census.

Tableau 8

Comparaison de l'EAM(N) produite par la méthode X-11-ARMMI (facteurs saisonniers courants) appliquée à un modèle additif et de l'EAM(N) produite par la même méthode appliquée à un modèle multiplicatif en périodes de récession et de non-récession

Séries	Période de récession (N = 24)		Période de non-récession (N = 60)	
	Modèle additif X-11-ARIMA facteurs saisonniers courants	Modèle multiplicatif X-11-ARMMI facteurs saisonniers courants	Modèle additif X-11-ARIMA facteurs saisonniers courants	Modèle multiplicatif X-11-ARMMI facteurs saisonniers courants
Chômage	Hommes, 25 ans et plus 1.25	1.25	Hommes, 25 ans et plus 1.15	1.15
	Femmes, 25 ans et plus 1.14	1.14	Femmes, 25 ans et plus 0.88	0.88
	Hommes, 15 à 24 1.23	1.23	Hommes, 15 à 24 1.05	1.05
	Femmes, 15 à 24 0.93	0.93	Femmes, 15 à 24 0.85	0.85
Emploi	Hommes, 25 ans et plus 1.25	1.25	Hommes, 25 ans et plus 1.25	1.25
	Femmes, 25 ans et plus 1.00	1.00	Femmes, 25 ans et plus 1.00	1.00
	Hommes, 15 à 24 0.80	0.80	Hommes, 15 à 24 0.80	0.80
	Femmes, 15 à 24 1.14	1.14	Femmes, 15 à 24 1.17	1.17

- 4) La méthode X-11 avec facteurs saisonniers prévus une année à l'avance est celle qui produit les estimations les moins précises pour toutes les séries et dans toutes les situations étudiées.
- 5) En comparant les EAM produites par la méthode X-11-ARMMI (facteurs saisonniers courants) en période de récession et en période de non-récession, nous constatons qu'elles sont à peu près du même ordre, sauf en ce qui a trait à la série du chômage chez les hommes de 25 ans et plus, pour laquelle les révisions sont beaucoup plus fortes en période de récession. Ce cas d'exception s'explique par les brusques variations saisonnières auxquelles cette série est soumise à cause des changements profonds que subit sa composition. Les révisions sont plus importantes à cause principalement de ces modifications.

Par ailleurs, la méthode X-11 avec facteurs saisonniers courants entraîne, pour la plupart des séries, des révisions très différentes selon qu'il s'agit d'une période de récession ou d'une période de non-récession. Cette divergence indique que les révisions tiennent surtout au fait que les filtres aux extrémités ne permettent pas d'estimer avec précision les ordres de grandeur, très volatiles, observés en période de récession. Il n'y a qu'une série pour laquelle les deux méthodes les plus efficaces entraînent des révisions beaucoup plus fortes en période de non-récession; c'est celle du chômage chez les femmes de 15 à 24 ans. Cela s'explique par le comportement particulier de cette série au cours de la période observée, lequel est caractérisé par de fortes augmentations annuelles (environ 15% entre 1966 et 1973 et 8.5% entre 1973 et 1980), qui masquent l'effet du cycle économique, et par une composante saisonnière indépendante de ce cycle.

Tableau 6

Comparaison de l'EAM(N) produite par la méthode X-11-ARMMI avec facteurs saisonniers courants et des EAM(N) produites par trois autres méthodes pour la désaisonnalisation des séries de composition additive sur l'emploi et le chômage en période de récession (N = 24)					
Séries			Chômage		
X-11	X-11-ARMA	X-11-ARMMI	X-11	X-11-ARMA	X-11-ARMMI
facteurs saisonniers courants	facteurs implicites prévus	facteurs saisonniers courants	facteurs saisonniers courants	facteurs implicites prévus	facteurs saisonniers courants
Hommes, 25 ans et plus	1.18	1.29	Hommes, 25 ans et plus	1.75	1.38
Femmes, 25 ans et plus	1.16	1.49	Hommes, 15 à 24	1.70	1.84
Hommes, 15 à 24	1.21	1.48	Hommes, 25 ans et plus	1.44	2.08
Femmes, 15 à 24	1.33	1.74	Femmes, 25 ans et plus	1.26	1.65
Hommes, 25 ans et plus	1.02	1.05	Hommes, 15 à 24	1.34	2.05
Hommes, 15 à 24	1.50	1.50	Femmes, 15 à 24		

Tableau 7

Comparaison de l'EAM(N) produites par la méthode X-11-ARMMI avec facteurs saisonniers courants et des EAM(N) produites par trois autres méthodes pour la désaisonnalisation des séries de composition additive sur l'emploi et le chômage en période de non-récession (N = 60)					
Séries			Chômage		
X-11	X-11-ARIMA	X-11-ARMMI	X-11	X-11-ARMMI	X-11-ARMMI
facteurs saisonniers courants	facteurs saisonniers prévus	facteurs saisonniers courants	facteurs saisonniers courants	facteurs saisonniers courants	facteurs saisonniers courants
Hommes, 25 ans et plus	1.31	1.65	Hommes, 25 ans et plus	1.31	1.88
Femmes, 25 ans et plus	1.20	1.59	Hommes, 15 à 24	1.71	1.89
Hommes, 15 à 24	1.22	1.57	Hommes, 25 ans et plus	1.26	1.54
Femmes, 15 à 24	1.05	1.20	Femmes, 25 ans et plus	1.30	1.55
Hommes, 25 ans et plus	1.16	1.24	Hommes, 15 à 24	2.16	
Hommes, 15 à 24	1.10	1.27			
Femmes, 15 à 24	1.22	1.31			
Hommes, 25 ans et plus	1.41	1.68			
Femmes, 15 à 24					

Afin de vérifier ces affirmations, nous avons désaisonnalisé les huit séries sur la population active analysées précédemment suivant un modèle additif. Les résultats de cet exercice confirment ceux qui ont été obtenus au moyen d'un modèle multiplicatif. En effet, la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne les plus faibles révisions; viennent ensuite les méthodes X-11 avec facteurs saisonniers courants et X-11-ARMMI avec facteurs prévus une année à l'avance. La méthode X-11 avec facteurs prévus une année à l'avance est celle qui produit les estimations les moins précises. Il importe de souligner que, dans la désaisonnalisation de séries de modèle additif, les facteurs sont des *facteurs implicites* en ce sens qu'ils découlent du quotient de la série d'origine par la série désaisonnalisée.

Les tableaux 6 et 7 comparent l'importance relative des révisions produites par chacune des trois méthodes moins efficaces et de celles produites par la méthode X-11-ARMMI avec facteurs saisonniers courants pour les périodes de récession et de non-récession respectivement. Toutes les valeurs des tableaux sont supérieures à l'unité, ce qui indique qu'aucune des trois méthodes n'entraîne des révisions inférieures à celles produites par la méthode X-11-ARMMI avec facteurs saisonniers courants. Comme celle-ci s'avère la plus efficace tant dans la version additive que dans la version multiplicative, nous allons chercher à savoir, à l'aide de chaque série, lequel des deux modèles de décomposition produit les plus faibles révisions.

En ce qui a trait aux deux séries qui influent le plus sur le taux de chômage (c'est-à-dire, le chômage chez les hommes de 25 ans et plus et l'emploi chez les hommes de 25 ans et plus), les données du tableau 8 indiquent que l'utilisation d'un modèle multiplicatif est préférable en période de récession comme en période de non-récession. Ces données confirment pour la plupart les modèles de décomposition qu'a choisis Statistique Canada en se fondant sur les tests de modèles (Morrison 1975; Higginson 1977). Seule la série de l'emploi chez les hommes de 15 à 24 ans semble faire exception à la règle puisque, dans ce cas, un modèle additif serait plus efficace. Mais comme les révisions sont déjà très faibles, un changement de modèle n'apporterait rien de plus. Les EAM, qui sont de 0.41 (période de récession) et de 0.39 (période de non-récession) avec un modèle multiplicatif, sont réduites à 0.33 et à 0.31 respectivement avec un modèle additif.

Enfin, les données du tableau 8 nous permettent de constater qu'un modèle multiplicatif conviendrait mieux qu'un modèle additif, en période de récession, pour la série du chômage chez les femmes de 25 ans et plus.

4. CONCLUSIONS

Les résultats présentés dans les sections 2 et 3 peuvent se résumer comme suit:

- 1) En période de récession comme en période de non-récession, la méthode X-11-ARMMI avec facteurs saisonniers courants entraîne les plus faibles révisions pour chaque série, peu importe si la série est de composition additive ou multiplicative.
- 2) Les ratios entre l'EAM produite par la méthode X-11-ARMMI (facteurs saisonniers courants) appliquée à un modèle additif et l'EAM produite par la même méthode appliquée à un modèle multiplicatif indiquent clairement que les deux séries qui influent le plus sur le taux de chômage (chômage chez les hommes de 25 ans et plus et emploi chez les hommes de 25 ans et plus) sont de composition multiplicative en période de récession comme en période de non-récession.
- 3) En période de récession, les méthodes X-11-ARMMI avec facteurs prévus une année à l'avance et X-11 avec facteurs saisonniers courants produisent des EAM identiques pour les séries de l'emploi et du chômage chez les hommes de 25 ans et plus. Pour ce qui a trait aux six autres séries toutefois, la méthode X-11 avec facteurs saisonniers courants s'avère la plus efficace après la X-11-ARMMI (facteurs saisonniers courants).

Tableau 4
Comparaison de l'EAM(N) produite par la méthode X-11-ARMMI avec facteurs saisonniers courants et des EAM(N) produites par trois autres méthodes pour la désaisonnalisation des séries de composition multiplicative sur l'emploi et le chômage en période de non-récession ($N = 60$)

Séries		X-11		X-11-ARMMI		X-11-ARMMI		X-11-ARMMI	
facteurs saisonniers courants		prévus		facteurs saisonniers courants		facteurs saisonniers courants		facteurs saisonniers courants	
X-11-ARMMI		X-11-ARMMI		X-11-ARMMI		X-11-ARMMI		X-11-ARMMI	
(1) ^a		(2) ^b		(3) ^c					
Hommes, 25 ans et plus	1.26	1.62	1.59	1.99	1.93	2.01	1.65	1.25	1.48
Femmes, 25 ans et plus	1.31	1.50	1.43	1.62	1.48	1.77	1.86	1.17	1.48
Hommes, 15 à 24	1.35	1.61	1.34	1.62	1.48	1.77	1.86	1.18	1.48
Femmes, 15 à 24	1.49	1.50	1.43	1.62	1.48	1.77	1.86	1.18	1.48
Hommes, 25 ans et plus	1.25	1.50	1.43	1.62	1.48	1.77	1.86	1.17	1.48
Femmes, 25 ans et plus	1.17	1.43	1.34	1.62	1.48	1.77	1.86	1.17	1.48
Hommes, 15 à 24	1.18	1.49	1.34	1.62	1.48	1.77	1.86	1.18	1.48
Femmes, 15 à 24	1.14	1.58	1.34	1.62	1.48	1.77	1.86	1.14	1.48

^a La colonne (1) est égale au quotient des colonnes (2) et (1) du tableau 3.
^b La colonne (2) est égale au quotient des colonnes (3) et (1) du tableau 3.
^c La colonne (3) est égale au quotient des colonnes (4) et (1) du tableau 3.

Tableau 5

Comparaison des EAM(N) des facteurs saisonniers courants des méthodes X-11-ARMMI et X-11 appliquées à des séries de composition multiplicative en période de récession et en période de non-récession

Séries		X-11-ARIMA		X-11	
facteurs courants		facteurs courants		facteurs courants	
période de récession ($N = 24$)		période de récession ($N = 24$)		période de récession ($N = 24$)	
(1) ^a		(2) ^b		(3) ^c	
Hommes, 25 ans et plus	1.42	1.59	1.22	1.35	0.60
Femmes, 25 ans et plus	1.05	1.07	1.22	1.35	0.60
Hommes, 15 à 24	1.09	1.35	1.22	1.35	0.60
Femmes, 15 à 24	0.67	1.35	1.22	1.35	0.60
Hommes, 25 ans et plus	1.00	1.20	1.07	1.27	1.54
Femmes, 25 ans et plus	1.00	1.20	1.07	1.27	1.54
Hommes, 15 à 24	1.05	1.27	1.07	1.27	1.54
Femmes, 15 à 24	1.16	1.54	1.07	1.27	1.54

^a La colonne (1) est égale au quotient de la colonne (1) du tableau 1 par la colonne (1) du tableau 3.
^b La colonne (2) est égale au quotient de la colonne (2) du tableau 1 par la colonne (2) du tableau 3.
^c La colonne (3) est égale au quotient de la colonne (3) du tableau 1 par la colonne (3) du tableau 3.

Le tableau 4 présente une comparaison de l'efficacité relative de la méthode X-11-ARMMI avec facteurs saisonniers courants et de celle des trois autres méthodes pour les années de non-récession. On constate que les chiffres de la première colonne, sauf pour ce qui a trait à une série, sont inférieurs à ceux de la première colonne du tableau 2. On peut en déduire que l'utilisation des extrapolations ARMMI permet de réaliser des gains en pourcentage en-core plus élevés en période de récession.

Enfin, le tableau 5 donne une comparaison de l'importance des révisions en période de récession et en période de non-récession pour les deux méthodes les plus efficaces. Les résultats indiquent que la méthode X-11-ARMMI avec facteurs saisonniers courants, qui est la méthode de désaisonnalisation officielle de Statistique Canada, produit des EAM inférieures à celles produites par la méthode X-11 avec facteurs saisonniers courants. La plupart des ratios de la première colonne sont très près de l'unité; nous pouvons en conclure que l'utilisation des extrapolations ARMMI entraîne des révisions du même ordre en période de récession comme en période de non-récession. Lorsqu'on applique la méthode X-11 avec facteurs saisonniers courants, les révisions sont sensiblement plus élevées en période de récession qu'en période "normale" pour la plupart des séries. Cette divergence tient au fait que les variations rapides de l'ordre de grandeur de la série, provoquées par les nouvelles observations relatives aux années de récession, ne sont pas estimées avec autant de précision par les filtres aux extrêmes. De fait, des mouvements progressifs et une partie de l'augmentation de l'ordre de grandeur de la série sont imputés à la composante saisonnière.

La seule série qui fait exception est celle du chômage chez les femmes de 15 à 24 ans; les deux méthodes entraînent dans ce cas des révisions plus faibles en période de difficultés économiques. Cela s'explique par le comportement particulier de cette série au cours de la période analysée, lequel est caractérisé par de fortes augmentations annuelles (environ 15% entre 1966 et 1973 et 8,5% entre 1973 et 1980) et une composante saisonnière additive, indépendante du cycle économique (c'est-à-dire que la variation de l'ordre de grandeur de la série exprimeait plus un changement de comportement des jeunes femmes qu'un effet du cycle économique). La série du chômage chez les hommes de 25 ans et plus constitue un autre cas particulier. En dépit d'extrapolations ARMMI, les années de récession présentent en effet des révisions beaucoup plus importantes que celles des années de non-récession, comme en fait foi le rapport de 1.42. Cet écart appréciable entre les révisions des deux périodes s'explique par les très fortes variations que subit la composition du mouvement saisonnier de cette série en période de récession. Sans extrapolations ARMMI, l'écart entre les révisions des deux périodes est encore plus grand (ratio de 1.59) puisque, outre les variations de la composition du mouvement saisonnier, les estimations saisonnières peu précises produites pour les années de récession contribuent à accroître cet écart.

3. DÉSaisonnalisation courante des séries de composition additive ou multiplicative en périodes de récession ET DE NON-RÉCESSION

On soutient souvent que le modèle de décomposition additif convient mieux que le modèle de décomposition multiplicatif en période de récession, si l'on vise à minimiser les révisions. Deux grandes raisons sont invoquées à l'appui de cette affirmation: (1) dans un modèle additif, les composantes de la série chronologique sont supposées indépendantes; ainsi, l'effet saisonnier ne se trouve pas influencé par le niveau de la tendance-cycle, contrairement à ce qui se passe dans un modèle multiplicatif; (2) les filtres aux extrêmes sont trop rigides pour permettre d'estimer avec précision un mouvement saisonnier rapide.

Comparaison de l'EAM(N) produites par la méthode X-11-ARMMI avec facteurs saisonniers courants et des EAM(N) produites par trois autres méthodes pour la désaisonnalisation des séries de composition multiplicative sur l'emploi et le chômage en période de récession ($N = 24$)

Séries		
X-11	X-11-ARMMI	X-11-ARMMI
Facteurs saisonniers prévus	Facteurs saisonniers prévus	Facteurs saisonniers prévus
X-11-ARMMI	X-11-ARMMI	X-11-ARMMI
Facteurs saisonniers courants	Facteurs saisonniers courants	Facteurs saisonniers courants
(3) ^c	(2) ^b	(1) ^a

a La colonne (1) est égale au quotient des colonnes (2) et (1) du tableau 1.
 b La colonne (2) est égale au quotient des colonnes (3) et (1) du tableau 1.
 c La colonne (3) est égale au quotient des colonnes (4) et (1) du tableau 1.

Erreurs absolues moyennes (EAM(N)) des facteurs saisonniers des méthodes X-11-ARMMI et X-11 durant les années de non-récession^a (N = 60)

Séries	(1) X-11-ARMMI	(2) X-11	(3) X-11-ARMMI	(4) X-11
	Facteurs saisonniers courants		Facteurs saisonniers prévus une année à l'avance	

Hommes, 25 ans et plus

Femmes, 25 ans et plus

HOMMES, 13 & 24 ANS

RECEIVED

Emploio!

Hommes, 25 ans et plus

Femmes, 25 ans et plus

HOMMES, 15 à 24 ans

FEMINISM, 1924 AND

* De janvier 1971 à décembre 1977, à l'exclusion des périodes de récession définies à la note (a) du tableau 1.

Le tableau 1 donne l'erreur absolue moyenne (EAM) des facteurs saisonniers des méthodes X-11-ARMMI et X-11 utilisées pour la désaisonnalisation courante en période de récession. Il est clair d'après ce tableau que la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne les plus faibles révisions. Cela confirme les résultats des études théoriques qui ont été commentées précédemment et selon lesquelles l'utilisation des extrapolations ARMMI avec des facteurs saisonniers courants avait pour effet de réduire sensiblement les révisions attribuables aux filtres.

La méthode X-11 avec facteurs saisonniers courants produit les plus faibles EAM après celles de la X-11-ARMMI avec facteurs saisonniers courants dans six séries sur huit. Pour ce qui a trait aux deux autres séries (chômage-hommes de 25 ans et plus et emploi - hommes de 25 ans et plus), la X-11 avec facteurs saisonniers courants produit les mêmes EAM que la méthode X-11-ARMMI avec facteurs saisonniers prévus une année à l'avance. Enfin, la méthode X-11 avec facteurs saisonniers prévus une année à l'avance est celle qui produit les estimations les moins précises.

Le tableau 2 présente une comparaison de l'efficacité relative de la méthode X-11-ARMMI avec facteurs saisonniers courants et de celle des trois autres méthodes, en période de récession. Les rapports, tous supérieurs à l'unité, indiquent que la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne effectivement les plus faibles révisions. La période de non-récession s'étend de janvier 1971 à décembre 1977, hors les années de récession. Le tableau 3 donne les EAM des séries désaisonnalisées courantes par suite de l'application des quatre méthodes en période de non-récession.

Comme au tableau 1, la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne les plus faibles révisions dans toutes les séries puisque, comme il a été mentionné plus haut, elle réduit au minimum les révisions attribuables aux filtres. La méthode X-11 avec facteurs saisonniers courants produit les deuxièmes meilleurs résultats dans sept séries sur huit, ces résultats étant relativement comparables à ceux obtenus par la méthode X-11-ARMMI avec facteurs saisonniers prévus une année à l'avance. Enfin, la méthode X-11 avec facteurs saisonniers prévus une année à l'avance est celle qui entraîne les révisions les plus importantes.

Tableau 1

Erreurs absolues moyennes (EAM(N)) des facteurs saisonniers des méthodes X-11-ARMMI et X-11 durant les années de récession^a (N = 24)

Séries	X-11-ARMMI facteurs saisonniers courants	X-11	X-11-ARMMI prévus une année à l'avance
X-11	(1)	(2)	(3)
(4)			
Chômage	1.95	2.75	2.74
Hommes, 25 ans et plus	1.94	2.94	3.43
Femmes, 25 ans et plus	2.16	3.02	3.49
Hommes, 15 à 24 ans	1.25	1.73	2.48
Femmes, 15 à 24 ans	0.08	0.12	0.12
Hommes, 25 ans et plus	0.23	0.29	0.33
Femmes, 25 ans et plus	0.41	0.53	0.66
Hommes, 15 à 24 ans	0.50	0.70	0.81
Femmes, 15 à 24 ans			0.97

^a D'août 1974 à juillet 1975 et de juin 1976 à mai 1977.

années pour produire une série désaisonnalisée. Par conséquent, 1971 est la première année pour laquelle nous pouvons calculer des mesures de la révision totale. L'année 1977 est la dernière année complète pour laquelle une EAM a été calculée. Les sept années se prêtant ainsi au calcul des révisions se caractérisent par deux années de récession et cinq années de non-récession.

La période de récession englobe les données d'août 1974 à juillet 1975 et de juin 1976 à mai 1977. Ces deux années sont considérées comme des années de récession parce qu'on y a observé une forte augmentation (supérieure à 25%) du nombre annuel de chômeurs imputable surtout au grand nombre de mises à pied.

Un autre aspect important dont les auteurs ont tenu compte a trait au genre de modèle de décomposition utilisé pour la désaisonnalisation de chaque série. La variante X-11 et le programme X-11-ARMMI produisent aussi bien une version additive qu'une version multiplicative. Il n'y a aucune raison théorique qui justifie la supériorité d'un modèle par rapport à l'autre. Les deux reposent sur des hypothèses différentes concernant le mécanisme de production de la composante saisonnière.

Dans un modèle additif, les éléments d'une série chronologique (tendance-cycle, variations saisonnières et aléas) sont supposés être indépendants. L'effet saisonnier n'est donc pas influencé par le niveau d'activité économique, lequel est déterminé par les étapes du cycle économique.

Dans un modèle multiplicatif, par ailleurs, l'effet saisonnier est proportionnel à la tendance-cycle. Si les facteurs saisonniers sont constants, cette relation signifie que plus l'ordre de grandeur de la série désaisonnalisée est élevé, plus l'effet saisonnier est prononcé.

Le choix du modèle de décomposition n'est pas déterminant pour l'estimation des valeurs désaisonnalisées (finales) puisque, la plupart du temps, les chiffres diffèrent peu d'un modèle à l'autre. Cependant, la question prend une toute autre importance lorsqu'il s'agit d'estimer la composante saisonnière des premières et des dernières années d'une série, surtout une série dont la tendance-cycle croît rapidement. Les filtres asymétriques utilisés pour l'estimation des valeurs extrêmes d'une série, particulièrement ceux de la méthode X-11, introduisent d'importantes erreurs systématiques lorsque les estimations saisonnières fluctuent rapidement (Dagum 1978). En fait, si le modèle de décomposition considéré est un modèle multiplicatif caractérisé par un mouvement saisonnier plutôt stable, la désaisonnalisation d'une série de composition additive produira des estimations saisonnières qui sembleront varier en fonction de la tendance-cycle. Réciproquement, s'il s'agit d'un modèle de décomposition additif caractérisé par un mouvement saisonnier stable, la désaisonnalisation d'une série de composition multiplicative produira des facteurs saisonniers dont le mouvement semblera imprévisible.

Du point de vue de la désaisonnalisation, il est donc préférable de choisir le modèle de décomposition qui produit les estimations saisonnières les plus stables. Les tests mis au point par Morry (1975) et Hinginson (1977) ont été appliqués aux huit séries pour déterminer quel modèle de décomposition convient le mieux.

Ces tests ont révélé que deux séries seulement (chômage chez les fermes de 25 ans et plus et celles de 15 à 24 ans) étaient de composition additive tandis que les autres étaient de composition multiplicative.

Dans le présent exposé, toutefois, nous analysons les révisions absolues moyennes en fonction des deux possibilités, c'est-à-dire une série de composition multiplicative ou une série de composition additive. Nous appliquons les deux genres de modèles de décomposition à des données couvrant à la fois des périodes de récession et des périodes de non-récession afin de déterminer lequel de ces modèles, du point de vue de la révision, est le plus sensible aux variations subtiles d'ordre de grandeur.

Les chiffres figurant dans les tableaux qui suivent ont été obtenus par la désaisonnalisation de séries de composition additive sont analysés à la section 3.

Les conclusions de ces deux études théoriques confirment les résultats présentés dans de nombreux travaux théoriques et empiriques (voir, par exemple, Dagum 1978; Dagum et Morry 1982; Kuiper 1978, 1981; Pierce 1980; Kenny et Durbin 1982; McKenzie 1982; Wallis 1982; Pierce et McKenzie 1985; Otto 1985).

Nous allons maintenant comparer l'efficacité de la méthode X-11-ARMMI utilisée avec des facteurs saisonniers courants et celle de trois autres méthodes de désaisonnalisation en périodes de récession et de non-récession. La méthode la plus efficace sera celle qui entraîne les plus faibles révisions.

2. 1 Comparaison de quatre méthodes de désaisonnalisation courante des séries sur la population active

Quatre méthodes de désaisonnalisation sont généralement utilisées pour produire des valeurs désaisonnalisées courantes; ce sont:

- 1) la méthode X-11-ARMMI avec facteurs saisonniers courants;
- 2) la méthode X-11 avec facteurs saisonniers courants;
- 3) la méthode X-11-ARMMI avec facteurs saisonniers prévus une année à l'avance;
- 4) la méthode X-11 avec facteurs saisonniers prévus une année à l'avance.

La mesure de la révision qui sert, en l'occurrence, à évaluer ces quatre méthodes est l'erreur absolue moyenne (EAM) des facteurs saisonniers pour la désaisonnalisation courante:

$$MAE(N) = \frac{\sum_{t=1}^N |S_t^c - S_t^f|}{N} \quad (1)$$

Dans l'équation ci-dessus, N représente le nombre d'éléments inclus dans la moyenne et S_t^c désigne la valeur courante du facteur saisonnier, lequel peut être un facteur saisonnier courant ou un facteur saisonnier prévu une année à l'avance (t un ou l'autre utilisé avec les méthodes X-11 ou X-11-ARMMI). S_t^f désigne la valeur "finale" du facteur saisonnier, c'est-à-dire que cette valeur ne variera pas beaucoup lorsque de nouvelles données s'ajouteront à la série. Pour la X-11 comme pour la X-11-ARMMI, la valeur courante d'un facteur saisonnier devient finale lorsque les observations d'au moins trois années sont ajoutées à la série (Young 1968; Wallis 1974).

Dans le présent exposé, on examine les révisions des facteurs saisonniers (ou des facteurs saisonniers implicites s'il s'agit du modèle additif) plutôt que celles des estimations désaisonnalisées, et ce pour plusieurs raisons. Premièrement, l'utilisation des facteurs saisonniers permet d'avoir une idée de l'importance des révisions par rapport à l'ordre de grandeur de la série (puisqu'il s'agit d'un pourcentage); deuxièmement, elle uniformise la valeur des révisions à l'intérieur des séries dont l'ordre de grandeur est soumis à de fortes variations (par exemple, les séries sur le chômage); troisièmement, elle permet de faire des recoupements entre les séries.

Contrairement à une étude antérieure faite par les mêmes auteurs (Dagum et Morry 1982), le présent exposé ne tient pas compte des révisions des variations mensuelles des données désaisonnalisées puisque ces révisions ne présentent pas un très grand intérêt lorsqu'on analyse des données sur la population active (Statistique Canada, par exemple, ne publie pas de révisions annuelles du taux de croissance pour ces séries). Cette étude porte donc sur les révisions de l'ordre de grandeur plutôt que sur les révisions de ses variations.

Les huit séries canadiennes sur l'emploi et le chômage analysées ici débutent en janvier 1966 et se terminent en octobre 1982. Pour pouvoir utiliser l'option des extrapolations ARMMI du programme X-11-ARMMI, il faut disposer de données remontant à au moins cinq

La présente étude vise principalement à vérifier si la méthode X-11-ARMMI utilisée avec des facteurs saisonniers courants entraîne toujours les plus faibles révisions en période de récession, comparativement à trois autres méthodes de désaisonnalisation.

Dans la section 2, on présente les erreurs absolues moyennes (EAM) des facteurs saisonniers courants et des facteurs saisonniers prévus une année à l'avance pour huit séries sur la population active du Canada désaisonnalisées au moyen des méthodes X-11-ARMMI et X-11 avec modèle de décomposition multiplicatif. On analyse les facteurs prévus une année à l'avance plutôt que ceux prévus six mois à l'avance parce que de nombreux organismes statistiques publiques les utilisent. En outre, les EAM des facteurs prévus six mois à l'avance se situent entre celles des facteurs courants et celles des facteurs prévus une année à l'avance. Dans la section 3, on calcule les erreurs absolues moyennes engendrées par les quatre méthodes de désaisonnalisation courante appliquées à un modèle additif et on les compare à celles obtenues avec un modèle multiplicatif.

Enfin, la section 4 contient les conclusions de l'étude.

2. RÉVISION DES FACTEURS SAISONNIERS SERVANT À LA DÉSAISON- NALISATION COURANTE DES SÉRIES SUR LA POPULATION ACTIVE EN PÉRIODES DE RÉCESSION ET DE NON-RÉCESSION

La plupart des méthodes de désaisonnalisation appliquées par les organismes statistiques publics sont fondées sur les filtres de lissage linéaires, appelés couramment moyennes mobiles. De par la nature même de ces méthodes, les estimations des observations des dernières années sont moins précises que celles des observations centrales à cause de l'asymétrie des filtres appliqués aux extrémités de la série. Les méthodes de désaisonnalisation les plus connues sont la variante X-11 de la Census Method II, mise au point par Shiskin, Young et Musgrave (1967), et la X-11-ARMMI, mise au point par Dagum (1980). La X-11-ARMMI est une version modifiée de la variante X-11 et comporte essentiellement deux étapes. Elle permet premièrement d'extrapoler la série originale au moyen de modèles ARMMI (modèles autorégressifs à moyennes mobiles intégrés) du type Box et Jenkins (1970) et de désaison-naliser la série ainsi prolongée au moyen d'un ensemble de moyennes mobiles qui sont une combinaison des filtres saisonniers de la variante X-11 et des filtres d'extrapolation de l'ARMMI. Les filtres de désaisonnalisation de la X-11-ARMMI et de la X-11 ne sont donc pas les mêmes pour ce qui a trait aux observations des dernières années. En ce qui concerne les observations centrales toutefois, les deux méthodes prévoient l'application du même filtre symétrique. Lorsque le modèle ARMMI n'est pas utilisé, la X-11-ARMMI équivaut à la méthode X-11.

A mesure que des données s'ajoutent, on révisé l'estimation désaisonnalisée se rapportant à une période donnée jusqu'à ce qu'il y ait un écart de trois ans entre celle-ci et la fin de la série et que les filtres symétriques puissent s'appliquer. L'estimation devient alors pratiquement fixe et est considérée comme une estimation désaisonnalisée finale. La différence entre la toute première estimation et l'estimation désaisonnalisée finale est définie comme la révision totale. L'application des méthodes X-11-ARMMI et X-11 entraîne une révision des valeurs désaisonnalisées courantes à cause des différences qui surgissent entre les filtres de lissage linéaires appliqués aux mêmes observations à mesure que de nouvelles données s'ajoutent et à cause des modifications que les nouvelles observations introduisent dans une série. Il serait souhaitable de supprimer ou de réduire au minimum les révisions causées par le premier facteur.

Des études théoriques réalisées par l'une des auteures (Dagum 1982a, 1982b) ont montré qu'il était possible de réduire sensiblement ces révisions en prolongeant la série d'origine avec des valeurs extrapolées ARMMI (c'est-à-dire, en appliquant la X-11-ARMMI) et en utilisant des facteurs saisonniers courants au lieu de facteurs saisonniers prévus une année à l'avance.

Désaisonnalisation des séries pour la population active en périodes de récession et de non-récession

ESTELA BEE DAGUM et MARILETTA MORRY¹

RÉSUMÉ

Les auteurs analysent les révisions de huit séries sur la population active désaisonnalisées en périodes de récession et de non-récession. Elles utilisent à cette fin les méthodes de désaisonnalisation X-11 et X-11-ARMMI appliquées avec des facteurs saisonniers courants ou avec des facteurs saisonniers prévus. La désaisonnalisation se fait suivant un modèle de décomposition additif et un modèle de décomposition multiplicatif. Les résultats indiquent que la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne les révisions les plus faibles tant en période de récession qu'en période de non-récession, et ce peu importe le modèle de décomposition utilisé.

MOTS CLÉS: X-11; X-11-ARMMI; désaisonnalisation courante; récession/non-récession.

1. INTRODUCTION

Le mouvement saisonnier de certaines séries sur la population active peut connaître des fluctuations soudaines à cause des changements majeurs que subit la composition de ces séries au cours des diverses étapes du cycle économique. La série sur le nombre total de chômeurs en est un exemple typique. Durant les années relativement prospères, cette série comprend surtout les personnes qui changent d'emploi, les nouveaux venus sur le marché du travail, les travailleurs du secteur primaire (agriculture, forêt, pêche, piégeage, etc.) et de la construction (en hiver) et les étudiants en quête d'un emploi (en été). En période de récession, toutefois, le nombre de chômeurs augmente rapidement et les nouveaux chômeurs sont surtout des travailleurs réguliers des industries lourdes et des secteurs connexes caractérisés par des variations saisonnières moins fortes et des structures saisonnières de l'emploi différentes de celles observées en période *normale*. Ce phénomène s'est produit au Canada en 1981-1982, lorsque le total non désaisonnalisé des chômeurs est passé de 790,000 en août 1981 à 1,494,000 en décembre 1982, les nouveaux chômeurs provenant principalement du secteur manufacturier et du secteur des services.

Compte tenu des variations subites de la taille et de la composition de la population des chômeurs durant les périodes de crise, il y a lieu de se demander si la méthode d'estimation des facteurs saisonniers, qui est appliquée aux années où le chômage est faible et surtout frictionnel et saisonnier, peut aussi s'appliquer aux années où le chômage est élevé et alimenté en grande partie par des pertes d'emploi dans les secteurs secondaire et tertiaire.

Au terme d'une recherche empirique effectuée à Statistique Canada en 1974, on a réussi à désaisonnaliser les séries sur la population active au moyen de la méthode X-11-ARMMI avec facteurs saisonniers courants. Dans les sections qui suivent, nous désignons cette méthode de désaisonnalisation la méthode *officielle*. En 1980, le U.S. Bureau of Labor Statistics a adopté officiellement la méthode X-11-ARMMI avec facteurs saisonniers prévus six mois à l'avance. Cet organisme applique également la méthode X-11-ARMMI avec facteurs saisonniers courants pour calculer le taux de chômage qu'il publie mensuellement. Les facteurs saisonniers courants s'obtiennent en désaisonnalisant chaque mois toutes les données connues jusqu'à ce jour, tandis que les facteurs saisonniers prévus proviennent de données qui remontent habituellement à un an (à six mois pour ce qui est du Bureau of Labor Statistics).

¹ Estela Bee Dagum et Marietta Morry, Division des séries chronologiques - recherche et analyse, Statistique Canada, 13^e étage, Immeuble R.H. Coats, parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

$$\frac{n_{AO}}{N_A} = \frac{\lambda_3 + \left(\frac{C_1 N_{A1}}{C_4 \lambda_3 \lambda_4 N_{B1}} \right)}{\frac{N_A}{N_B} \lambda_1 + \frac{n_B}{N_B} \lambda_2}$$

et

$$\frac{n_{BO}}{N_B} = \frac{\lambda_4 + \left(\frac{C_4 N_{B1}}{C_1 \lambda_3 \lambda_4 N_{A1}} \right)}{\frac{N_A}{N_B} \lambda_1 + \frac{n_B}{N_B} \lambda_2}$$

En utilisant ces relations, on peut calculer le coût C^* de la façon suivante:

$$C^* = \frac{(\xi_1 + \xi_2) [\{ C_1 (1 + \alpha_1^*) (\Phi_1' + \alpha_1^* D_1')^2 \}^{1/2} + \alpha_1^* (C_4 q_1')^{1/2}]^2}{\frac{1}{(\Phi_1' + \alpha_1^* D_1')^2} \frac{n_A}{\alpha \alpha_1^* q_2} + \frac{n_B}{\alpha \alpha_1^* q_2}} \tag{A.4}$$

BIBLIOGRAPHIE

ARMSTRONG, B. (1979). Test for multiple frames sampling technique for agricultural survey: Nouveau-Brunswick, 1978. *Techniques d'enquête*, 5, 178-199.

BOSECKER, R. R., et FORD, B. L. (1976). Multiple frame estimation with stratified overlap domain. *American Statistical Association Proceedings of the Social Statistics Section*, 219-224.

HARTLEY, H. O. (1962). Multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 203-206.

HARTLEY, H. O. (1974). Multiple frame methodology and selected application. *Sankhya*, Series C, 36, 99-118.

LUND, R. E. (1968). Estimation in multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 282-288.

SERRURIER, D., and PHILLIPS, J. (1976). Double frame Ontario pilot hog surveys. *Survey Methodology*, 2, 138-170.

VOGEL, F. A. (1975). Surveys with overlapping frames, problems in application. *American Statistical Association Proceedings of the Social Statistics Section*, 695-699.

Définissons

$$\lambda_1 = (N^{a(1)} + N^{a(12)}\sigma_2^2 + p^2(N^{ab(1)} + N^{ab(12)}\sigma_2^2) \sigma_{ab(1)}^2,$$

$$\lambda_2 = q^2(N^{ab(1)} + N^{ab(12)}\sigma_2^2) \sigma_{ab(1)}^2,$$

$$\lambda_3 = (N^{a(1)} + N^{a(12)}\sigma_2^2 + p^2(N^{ab(1)} + N^{ab(12)}\sigma_2^2) \sigma_{ab(1)}^2,$$

$$\lambda_4 = q^2(N^{ab(1)} + N^{ab(12)}\sigma_2^2) \sigma_{ab(1)}^2.$$

Etant donné le p' de l'équation (A.1), les tailles optimales des échantillons seront

$$n_{2AO}^{N_A} = \gamma' \frac{(N^{a(1)} + N^{a(12)}\sigma_2^2 + p'^2(N^{ab(1)} + N^{ab(12)}\sigma_2^2) \sigma_{ab(1)}^2)}{C_A'}$$

$$\lambda_3 = \gamma' \frac{C_A'}{\lambda_3}$$

$$\frac{n_{2BO}^{N_B}}{N_B} = \gamma' \frac{C_B'}{q'^2(N^{ab(1)} + N^{ab(12)}\sigma_2^2) \sigma_{ab(1)}^2} = \gamma' \frac{C_B'}{\lambda_4'}$$

avec γ' déterminé en fonction de (18). À partir de cela, on obtient

$$\frac{n_{BO}}{n_{AO}} = \frac{N_B}{N_A} \left(\frac{C_1 N_{A1} \lambda_4}{C_4 N_{B1} \lambda_3} \right)^{\frac{1}{2}}. \tag{A.2}$$

De plus, les variances calculées à l'aide de (5) et (17) pour des tailles d'échantillons optimales peuvent s'écrire

$$V(X_{(1)}) = \frac{N_A}{N_B} \lambda_1 + \frac{n_A}{n_B} \lambda_2 \tag{A.3}$$

$$V(X_{(1)}^*) = \frac{N_A}{N_B} \lambda_3 + \frac{n_{AO}}{n_{BO}} \lambda_4.$$

En égalant les variances ci-dessus et en utilisant l'équation (A.2), on obtient l'expression pour n_{AO} et n_{BO} en termes de n_A et n_B de la façon suivante:

ANNEXE

La solution optimale pour p' qui minimise la variance (17) sous la contrainte de C^* selon les hypothèses d'un champ d'observation de 100% par la base A et l'égalité des variances est donnée par

$$p',2 = \frac{1 - \alpha}{\sigma_2^{a(1)}(\xi_1 + \xi_2)} \left\{ \frac{\sigma_2^{ab(1)}(\xi_3 + \xi_4)}{\alpha} \right.$$

avec

$$\vartheta' = \frac{C'_A}{C'_B}.$$

En se servant de ce que $N_{A1} = N_{a(1)} + N_{a(12)} + N_{ab(1)} + N_{ab(12)}$ et de ce que $N_{B1} = N_{ab(1)} + N_{ab(12)}$, on peut écrire ϑ' de la façon suivante

$$\vartheta' = \frac{C_1}{C_4} \alpha \frac{N_a(\xi_1 + \xi_2) + N_{ab}(\xi_3 + \xi_4)}{N_{ab}(\xi_3 + \xi_4)}$$

$$= \frac{C_1}{C_4} \alpha \left(\frac{1 - \alpha}{\alpha} \frac{\xi_1 + \xi_2}{\xi_3 + \xi_4} + 1 \right)$$

$$= \frac{\alpha}{1} \left(\frac{K}{\alpha} \frac{1}{\alpha_1^*} + 1 \right)$$

où

$$\alpha_1^* = \frac{\alpha}{\xi_3 + \xi_4} \frac{1 - \alpha}{\xi_1 + \xi_2}.$$

D'où

$$p',2 = \frac{1 - \alpha}{\xi_1 + \xi_2} \frac{\frac{\alpha}{1} \left(\frac{K}{\alpha} \frac{1}{\alpha_1^*} + 1 \right) - \alpha}{\xi_3 + \xi_4} \Phi'_1$$

$$= \frac{K \Phi'_1}{1 + \alpha_1^*(1 - K)}$$

(A.1)

La réduction des coûts attribuable à l'utilisation d'une enquête à bases multiples plutôt qu'à l'utilisation d'une enquête à base unique a été calculée pour un ensemble donné de valeurs des paramètres et les résultats figurent dans le tableau 2. D'après ce tableau, la réduction des coûts est considérable.

Gain en pourcentage au niveau des coûts attribuable à l'utilisation d'une enquête commune à bases multiples portant sur deux caractères plutôt qu'à l'utilisation d'enquêtes distinctes

Lorsque $g = 10$, $\Phi_1 = 0.25$, $\Phi_2 = 0.5$, $\Phi_3 = 1$, $\alpha = 0.5$.

Tableau 1

g_3	$\xi_1 = 0.2, \xi_2 = 0.2, \xi_3 = 0.4, \xi_4 = 0.2$						$\xi_1 = 0.2, \xi_2 = 0.4, \xi_3 = 0.2, \xi_4 = 0.4$					
	0.3	0.4	0.5	0.6	0.7	0.8	0.3	0.4	0.5	0.6	0.7	0.8
0.9	3.9	4.2	4.4	4.4	6.9	8.9	15.9	1.5	1.7	1.8	4.5	14.5
0.8	6.4	6.7	6.9	7.0	9.1	10.9	12.6	3.9	4.2	4.5	6.9	14.3
0.7	8.7	8.9	9.1	9.3	11.2	12.8	14.3	6.4	6.7	6.9	9.1	15.9
0.6	10.7	10.9	11.2	11.3	13.0	14.5	15.9	8.7	8.9	9.1	11.2	14.3
0.5	12.6	12.8	13.0	13.3	14.5	15.9	17.9	10.7	10.9	11.2	13.0	14.5
0.4	14.3	14.5	14.7	14.9	16.1	17.3	19.5	12.6	12.8	13.0	14.5	15.9
0.3	15.9	16.1	16.3	16.5	17.7	18.9	20.1	14.3	14.5	14.7	16.1	17.3
0.2	17.3	17.5	17.7	17.9	19.1	20.3	21.5	15.9	16.1	16.3	17.7	18.9
0.1	18.9	19.1	19.3	19.5	20.7	21.9	23.1	17.3	17.5	17.7	19.1	20.3

Tableau 2

Réduction des coûts pour des variances constantes
Lorsque $\Phi_1 = 0.25$, $\Phi_2 = 0.5$, $\Phi_3 = 1$, et $\xi_1 = 0.2$, $\xi_2 = 0.3$, $\xi_3 = 0.4$.

α	0.5	0.6	0.7	0.8	0.9	0.95
100	.227	.175	.132	.094	.059	.040
20	.304	.254	.200	.169	.127	.101
10	.367	.321	.279	.238	.193	.164
5	.462	.423	.387	.351	.308	.277
2	.661	.646	.634	.621	.599	.578
1	.876	.895	.918	.943	.971	.985

Avec toutes ces substitutions apportées à l'expression (21) $(C^* + C^{**})/C$ se simplifie comme suit:

$$\frac{C^* + C^{**}}{C} = \frac{T_2^2}{\Phi + \alpha_1^* p} \times \frac{\left\{ \frac{r\alpha + K}{K(1 - \alpha)} + 1 \right\} \left\{ \theta_1 \xi + \xi + \theta_3(1 - 2\xi) \right\}}{2\xi\theta_1 + (1 - \xi)\theta_3}$$

$$= \frac{T_1^2}{(\Phi + \alpha_1^* p) \left\{ \frac{r\alpha + K}{K(1 - \alpha)} + 1 \right\}} \times \frac{\theta_3 + \xi(2\theta_1 - \theta_3)}{\theta_3 + \xi(\theta_1 + 1 - 2\theta_3)}$$

$$= \frac{\theta_3 + \xi(2\theta_1 - \theta_3)}{\theta_3 + \xi(1 + \theta_1 - 2\theta_3)}$$

où $r = (n_a/n_b)$ optimal $= p/\alpha q$ de (10).

$$G = \frac{\xi(\theta_1 + \theta_3 - 1)}{\theta_3 + \xi(1 + \theta_1 - 2\theta_3)} \times 100.$$

L'équation des ξ_i ne semble pas être une hypothèse réaliste. La valeur de G a donc été calculée à l'aide de (19) pour des combinaisons réalistes et représentatives de paramètres; les résultats obtenus figurent dans le tableau 1.
Ce tableau révèle qu'il y a manifestement un gain attribuable à l'intégration d'enquêtes à bases multiples portant sur deux caractères par rapport à la réalisation d'enquêtes distinctes. Le gain augment à mesure qu'augmentent les valeurs de θ_1 et θ_3 .

4. COMPARISON D'ENQUÊTES À BASE MULTIPLES PORTANT SUR DEUX CARACTÈRES ET D'UNE ENQUÊTE À BASE UNIQUE

La comparaison d'une enquête à deux bases de sondage et des enquêtes à base unique pour l'étude de deux caractères présente de l'intérêt sur le plan pratique. Une étude analogue a été faite par Hartley (1962) pour le cas d'un seul caractère. La réduction relative des coûts calculée suivant une méthode analogue s'exprime

$$R = \left(1 + \frac{p}{\alpha q} \right) \left/ \left(1 + \frac{p}{\alpha q(1 + p)} \right) \right.$$

où p^2 is given by (16), $\theta = CA/C_B$ et $\alpha = N^{ab}/N_a$

ou

$$T_1 = \left\{ (\Phi'_1 + \alpha_1^* p_{12}) (1 + \alpha_1^*) \right\}^{1/2} + \alpha_1^* q' \sqrt{K}$$
$$T_2 = \left\{ (\Phi'_2 + \alpha_2^* p_{22}) (1 + \alpha_2^*) \right\}^{1/2} + \alpha_2^* q'' \sqrt{K}.$$

K peut être calculé de la façon suivante: à l'aide de la définition de C_A , C_B , Φ , et de $i) = 1, \dots, 6$ et de l'équation (A.1), on obtient

$$C_A = \frac{1}{C_B} \frac{K}{\varrho_1 \Phi_1 + \Phi_2 + \varrho_3 \Phi_3} = \varrho_3,$$

et donc

$$(22) \quad K = \varrho^{-1} \left\{ \alpha + (1 - \alpha) \frac{\varrho_1 \xi_1 + \xi_2 + \varrho_3 (1 - \xi_1 - \xi_2)}{\varrho_1 \xi_3 + \xi_4 + \varrho_3 (1 - \xi_3 - \xi_4)} \right\}.$$

On peut utiliser l'expression de l'équation (21) pour calculer le gain au niveau des coûts que produit une étude portant sur les deux caractères simultanément par rapport à des enquêtes individuelles indépendantes. Le gain G exprimé en pourcentage est donc donné par

$$G = \left(\frac{C}{C^* + C^{**}} - 1 \right) \times 100$$

Dans la comparaison des coûts ci-dessus, les coûts prévus, C , C^* et C^{**} n'incluent pas les frais généraux afférents à la réalisation d'une enquête combinée ou à la réalisation d'enquêtes distinctes. On s'attend toutefois que le total des frais généraux d'enquêtes distinctes soit beaucoup plus élevé que le total correspondant se rapportant à la réalisation d'une enquête combinée. Par conséquent, le gain réel attribuable à la réalisation d'enquêtes indépendantes sera supérieur au gain G exprimé en pourcentage, tel qu'il est défini plus haut.

L'expression (21) se simplifie de façon appréciable en fonction des hypothèses $\Phi'_1 = \Phi'_2 = \Phi$ et $\xi_1 = \xi_2 = \xi_3 = \xi_4 = \xi$ (disons).

À partir de (22) $\varrho = 1/K$ et de (16) puisque $\Phi'_1/\Phi'_2 = \Phi'_3/\Phi'_4$, le p^2 se simplifie comme suit:

$$p^2 = \frac{K(1 - \alpha)}{1 - k\alpha} \Phi.$$

En outre $\alpha_1^* = \alpha_2^* = \alpha/(1 - \alpha)$.

Par conséquent, à partir de (A.1)

$$p' = p'' = \left\{ \frac{K(1 - \alpha)\Phi}{1 - K\alpha} \right\}^{1/2}$$

D'où

$$T_1 = T_2 = \left\{ \Phi \frac{1 - K\alpha}{1 - \alpha} \right\}^{1/2} + \frac{1}{\alpha \sqrt{K}}.$$

où

$$\alpha_1^* = \frac{\alpha}{\xi_3 + \xi_4} \frac{1 - \alpha}{\xi_1 + \xi_2}$$

De la même façon, pour l'enquête distincte portant sur le deuxième caractère, le coût se calcule somme suit:

$$C^{**} = \frac{(1 - \xi_1) [\{ C_3(1 + \alpha_2^*)(\Phi_2' + \alpha_2^* d^{''2}) + \alpha_2^* d^{''2} \}^{1/2} + \alpha_2^*(C_6 q^{''2})^{1/2}]}{\frac{1}{(\Phi_2' + \alpha_2^* d^2)} + \frac{\alpha}{n_b} \left\{ \frac{1 - \alpha}{n_a} \right\}} \quad (19)$$

où

$$p^{''2} = \frac{K \Phi_2'}{1 + \alpha_2^*(1 - K)}, \alpha_2^* = \frac{1 - \alpha}{\alpha} \frac{1 - \xi_1}{1 - \xi_2}$$

Pour l'étude combinée portant sur les deux caractères, le coût total C dans le cas d'un champ d'observation de 100% de la base A est donné par (8).

Par conséquent

$$C = \frac{N_A}{n_A} [C_1(N^{a(1)} + N^{ab(1)}) + C_2(N^{a(12)} + N^{ab(12)}) + C_3(N^{a(2)} + N^{ab(2)})] + \frac{N_B}{n_B} [C_4 N^{ab(1)} + C_5 N^{ab(12)} + C_6 N^{ab(2)}]$$

En utilisant les hypothèses formulées au sujet des coûts (c'est-à-dire, en supposant que $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$) on obtient

$$C = C_2 n_A [(1 - \alpha) \{ \xi_1 \xi_1 + \xi_2 + \xi_3(1 - \xi_1 - \xi_2) \} + C_2 n_A [(1 - \alpha) \{ \xi_1 \xi_1 + \xi_2 + \xi_3(1 - \xi_1 - \xi_2) \} + \frac{K}{\alpha} \{ \xi_1 \xi_3 + \xi_4 + \xi_3(1 - \xi_3 - \xi_4) \}] \quad (20)$$

où $r = n_A/n_B$, $\xi_1 = C_1/C_2$ and $\xi_3 = C_3/C_2$.

Mais dans l'étude combinée portant sur les deux caractères (n_A/n_B) optimal = $p/\alpha q$ où p est donné par (16). Donc, la valeur du gain peut être obtenue à partir du ratio.

$$\frac{C}{C^* + C^{**}} = \frac{(\xi_1 + \xi_2) \xi_1 T_1^2}{(1 - \xi_1) \xi_3 T_2^2} + \frac{(\Phi_1' + \alpha_1^* d)}{(\Phi_3' + \alpha_3^* d)} \left\{ \frac{r \alpha + K}{r(1 - \alpha)} \right\} \{ \xi_1 \xi_1 + \xi_2 + \xi_3(1 - \xi_1 - \xi_2) \} + \{ \xi_1 \xi_3 + \xi_4 + \xi_3(1 - \xi_3 - \xi_4) \} \quad (21)$$

sont les tailles des échantillons aléatoires et $E(n_{A1}^*) = n_A N_{A1}/N_A$ et $E(n_{B1}^*) = n_B N_{B1}/N_B$. Dans ce cas, l'estimateur $Y^{(1)*}$ et sa variance sont donnés par:

$$Y^{(1)*} = (N^{a(1)} + N^{a(12)})Y^{(a(1), a(12))} + (N^{ab(1)} + N^{ab(12)})(p'Y^{(ab(1), ab(12))} + q'Y^{(ba(1), ba(12))}) + (N^{b(1)} + N^{b(12)})Y^{(b(1), b(12))}$$

où p' et q' sont les variables de pondération de sorte que $p' + q' = 1$ et où $Y^{(a(1), a(12))}$, $Y^{(ab(1), ab(12))}$, etc. sont les moyennes des échantillons tirés des domaines combinés respectifs, c'est-à-dire, par exemple, que $Y^{(a(1), a(12))}$ est la moyenne des unités de l'échantillon tiré des domaines $a(1)$ and $a(12)$.

$$V(Y^{(1)*}) = \frac{N_A}{N_A} N^{a(1)} \sigma_2^{a(1)} + N^{a(12)} \sigma_{(12)}^{a(12)} + (p'^2 \frac{N_A}{N_A} N^{ab(1)} \sigma_2^{ab(1)} + q'^2 \frac{N_B}{N_B} N^{ab(12)} \sigma_{(12)}^{ab(12)} + N^{ab(12)} \sigma_{(12)}^{ab(12)}) + \frac{N_B}{N_B} N^{b(1)} \sigma_2^{b(1)} + N^{b(12)} \sigma_{(12)}^{b(12)}.$$

(17)

Dans ce cas, la fonction de coût prend la forme

$$C = C_1 n_{A1}^* + C_4 n_{B1}^*$$

et le coût prévu est représenté par C^*

$$C^* = C_1 \frac{N_A}{N_A} N_{A1} + C_4 \frac{N_B}{N_B} N_{B1} = C_1' n_A + C_4' n_B \quad (18)$$

où $C_1' = C_1 N_{A1}/N_A$ et $C_4' = C_4 N_{B1}/N_B$.

Pour simplifier, nous supposons un champ d'observation de 100% par la base A et l'égalité des variances comme dans (15), et nous supposons que $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$. Suivant ces hypothèses, le coût C^* avec n_A et n_B qui minimisent la variance (17) se calcule comme suit (voir annexe pour le détail du calcul).

$$C^* = \frac{(\xi_1 + \xi_2) \{ C_1 (1 + \alpha^*) (\Phi_1' + \alpha_1^* d'^2) + \alpha_1^* (C_4 q'^2) \}^{1/2} + \alpha_1^* (C_4 q'^2) \}^{1/2}}{\left\{ \frac{1}{\alpha} - \frac{1}{\alpha} \right\} (\Phi_1' + \alpha_1^* d'^2) + \frac{n_A}{\alpha \alpha_1^* q'^2} + \frac{n_B}{\alpha \alpha_1^* q'^2}}$$

Supposons que

(15) $\sigma_2^{a(1)2} = \sigma_2^{a(12)2}, \sigma_2^{a(2)2} = \sigma_2^{a(21)2}, \sigma_2^{ab(1)2} = \sigma_2^{ab(12)2}, \sigma_2^{ab(2)2} = \sigma_2^{ab(21)2}.$

Ces hypothèses semblent plausibles étant donné qu'il est peu probable que la variabilité d'un caractère soit touchée par la présence ou l'absence de l'autre caractère. Alors, p^2 se ramène à

$$p^2 = \frac{\alpha}{\alpha} \left\{ \frac{\sigma_2^{a(1)}(N^{a(1)} + N^{a(2)}) + \sigma_2^{a(2)}(N^{a(2)} + N^{a(12)})}{\sigma_2^{ab(1)}(N^{ab(1)} + N^{ab(12)}) + \sigma_2^{ab(2)}(N^{ab(2)} + N^{ab(12)})} \right\}$$

ou enfin à

(16) $p^2 = \frac{(1 - \alpha)\Phi_2 \left\{ \Phi_3(\xi_1 + \xi_2) + (1 - \xi_1) \right\}}{(1 - \alpha)\Phi_2 \left\{ \Phi_4(\xi_3 + \xi_4) + (1 - \xi_3) \right\}}$

où

$$\Phi_1' = \frac{\sigma_2^{a(1)}}{\sigma_2^{ab(1)}}, \Phi_2' = \frac{\sigma_2^{a(2)}}{\sigma_2^{ab(2)}}, \Phi_3' = \frac{\sigma_2^{a(1)}}{\sigma_2^{ab(1)}}, \Phi_4' = \frac{\sigma_2^{a(2)}}{\sigma_2^{ab(2)}}.$$

et

$$\xi_1 = \frac{N^{a(1)}}{N^{a(12)}}, \xi_2 = \frac{N^{a(2)}}{N^{a(12)}}, \xi_3 = \frac{N^{ab(1)}}{N^{ab(12)}}, \xi_4 = \frac{N^{ab(2)}}{N^{ab(12)}}.$$

En se servant du fait que $N^{ab} = N^a, N^{a(2)} + N^{a(12)} = N^a - N^{a(1)}$, et $N^{ab(2)} + N^{ab(12)} = N^{ab} - N^{ab(1)}$, on peut voir que l'expression ci-dessus se ramène à la forme habituelle du cas à un seul caractère étant donné que $\xi_1 = \xi_3 = 1$ et $\xi_2 = \xi_4 = 0$. Il convient de noter que les variances dans des domaines ne sont en général pas connues étant donné que de telles valeurs sont fondées sur des renseignements antérieurs ou sur des valeurs estimées. L'op-

3. COMPARISON D'UNE ENQUÊTE PORTANT SUR PLUSIEURS
CARACTÈRES ET DES ENQUÊTES INDÉPENDANTES
PORTANT SUR UN SEUL CARACTÈRE DANS DES
SITUATIONS À BASES MULTIPLES

Les enquêtes portant sur plusieurs caractères sont conçues dans le but d'économiser les ressources disponibles, et on peut s'attendre qu'une enquête commune présente des avan-
tages supérieurs à ceux des enquêtes indépendantes portant sur un seul caractère, sur le plan
des coûts et de l'efficacité. Nous allons donc examiner l'importance des avantages (gain)
attribuables à l'utilisation d'une enquête commune à bases de sondage multiples.
Soit, dans une étude portant sur un seul caractère, disons $y^{(1)}$, des échantillons aléatoires
simples de taille n_A et n_B tirés des bases de sondage A et B respectivement. Nous supposons
ici que les seules bases disponibles sont les bases utilisées précédemment, et non pas les bases
réduites pour chaque caractère. Définitions N_{A1}, N_{B1}, n_{A1}^* , et n_{B1}^* comme les tailles de la
population et les tailles des échantillons portant sur le caractère $y^{(1)}$. Ici, n_{A1}^* and n_{B1}^*

Pour obtenir les p_i optimums, tout comme n_A et n_B optimums, il faut minimiser la fonction F sous réserve de l'existence de la fonction de coût prévu comme celle calculée dans (9). Les variables de pondération p_i et les tailles des échantillons sont obtenues comme suit à l'aide du multiplicateur de Lagrange:

$$(10) \quad \frac{p_1}{P_1} = \frac{p_2}{P_2} = \frac{p_3}{P_3} = \frac{p_4}{P_4} = \frac{n_B n_A}{P} = \frac{n_B N_A}{q} \quad (\text{disons}),$$

$$\frac{n_A}{N_A} = \gamma \frac{C_A}{K_5 + K_1 p_1^2 + K_2 p_2^2 + K_3 p_3^2 + K_4 p_4^2},$$

$$(11) \quad \frac{n_B}{N_B} = \gamma \frac{C_B}{K_6 + K_1 q_1^2 + K_2 q_2^2 + K_3 q_3^2 + K_4 q_4^2},$$

où γ est déterminé pour satisfaire au coût prévu et où

$$K_1 = N^{ab(1)} \sigma_2^{ab(1)}, K_2 = N^{ab(12)} \sigma_{(1)2}^{ab(12)},$$

$$K_3 = N^{ab(2)} \sigma_2^{ab(2)}, K_4 = N^{ab(12)} \sigma_{(2)2}^{ab(12)},$$

$$K_5 = N^{a(1)} \sigma_2^{a(1)} + N^{a(2)} \sigma_2^{a(2)} + N^{a(12)} \sigma_{(1)2}^{a(12)} + \sigma_{(2)2}^{a(12)},$$

$$(12) \quad K_6 = N^{b(1)} \sigma_2^{b(1)} + N^{b(2)} \sigma_2^{b(2)} + N^{b(12)} \sigma_{(1)2}^{b(12)} + \sigma_{(2)2}^{b(12)}.$$

À partir de (10) et (11), on obtient

$$(13) \quad \frac{q^2 N_B C_B}{K_6 + (K_1 + K_2 + K_3 + K_4) q^2} = \frac{p^2 N_A C_A}{K_5 + (K_1 + K_2 + K_3 + K_4) p^2}$$

Il s'agit d'une fonction bi-quadratique en p qui peut être résolue pour p . Les fractions de sondage optimums peuvent être calculées à l'aide de (11). Un des cas qui se présente souvent dans les sondages à bases multiples est celui où les bases englobent la population totale. Prenons le cas où la base A englobe la population totale, alors $N^{b(1)} = N^{b(2)} = N^{b(12)} = 0$.

Dans ce cas, (13) se ramène à

$$(14) \quad p^2 = \frac{\alpha}{K_5} \frac{\alpha - \alpha}{K_1 + K_2 + K_3 + K_4} \quad \text{and} \quad \alpha = \frac{C_A}{C_B} \quad \text{and} \quad \alpha = \frac{N_A}{N_B}.$$

Pareillement,

$$W(Y_2) = \left(\frac{n_A}{N_A} \right) \left(N^{a(2)} \sigma_2^{a(2)} + N^{a(12)} \sigma_2^{a(12)} + P_2^3 N^{ab(2)} \sigma_2^{ab(2)} \right) + \left\{ \frac{n_B}{N_B} \right\} \left\{ N^{b(2)} \sigma_2^{b(2)} + N^{b(12)} \sigma_2^{b(12)} + P_2^4 N^{ab(12)} \sigma_2^{ab(12)} + P_2^3 N^{ab(2)} \sigma_2^{ab(2)} + P_2^4 N^{ab(12)} \sigma_2^{ab(12)} \right\} \quad (6)$$

où $\sigma_2^{a(1)}$, $\sigma_2^{a(2)}$, etc. sont les variances pour les deux caractères dans les sous-domaines respectifs. Pour optimiser les p_i 's ($i = 1, 2, 3, 4$) dans une enquête commune, il faut minimiser une combinaison de variances individuelles sous réserve du maintien du coût fixe total de l'enquête combinée. Prenons la combinaison linéaire la plus simple possible

$$F = W(Y_1) + W(Y_2).$$

Dans le cas de l'enquête commune, une fonction de coût qui conviendrait être la suivante:

$$C' = C_1(n^{a(1)} + n^{ab(1)}) + C_2(n^{a(12)} + n^{ab(12)}) + C_3(n^{a(2)} + n^{ab(2)}) + C_4(n^{b(1)} + n^{ba(1)}) + C_5(n^{b(12)} + n^{ba(12)}) + C_6(n^{b(2)} + n^{ba(2)}) \quad (7)$$

où C_1 est le coût par unité dans le sous-domaine $a(1)$, $ab(1)$; C_2 dans le sous-domaine $a(2)$, $ab(2)$; C_3 dans le sous-domaine $a(12)$, $ab(12)$; C_4 dans le sous-domaine $a(2)$, $ab(2)$ de la base A . De même C_5 et C_6 sont les coûts par unité dans les sous-domaines de la base B . Dans la fonction de coût ci-haut, la taille des échantillons pris au hasard entre en jeu. Soit le coût prévu

$$C = E(C') = n_A(C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3) + n_B(C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6) \quad (8)$$

où

$$\Phi_1 = \frac{N^{a(1)} + N^{ab(1)}}{N^{a(1)} + N^{ab(1)} + N^{a(12)} + N^{ab(12)}}, \quad \Phi_2 = \frac{N^{a(2)} + N^{ab(2)}}{N^{a(2)} + N^{ab(2)} + N^{a(12)} + N^{ab(12)}}, \quad \Phi_3 = \frac{N^{a(12)} + N^{ab(12)}}{N^{a(12)} + N^{ab(12)} + N^{a(1)} + N^{ab(1)}}, \quad \Phi_4 = \frac{N^{a(12)} + N^{ab(12)}}{N^{a(12)} + N^{ab(12)} + N^{a(1)} + N^{ab(1)}}, \quad \Phi_5 = \frac{N^{b(12)} + N^{ba(12)}}{N^{b(12)} + N^{ba(12)} + N^{b(2)} + N^{ba(2)}}, \quad \Phi_6 = \frac{N^{b(12)} + N^{ba(12)}}{N^{b(12)} + N^{ba(12)} + N^{b(2)} + N^{ba(2)}}.$$

Où

$$C = n_A C_A + n_B C_B \quad (9)$$

où

$$C_A = C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3 \quad \text{et} \quad C_B = C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6.$$

Par conséquent,

$$Y^{(1)} = N^{a(1)}y^{a(1)} + N^{a(12)}y^{a(12)} + N^{ab(1)}(p_1y^{ab(1)} + q_1y^{ba(1)}) + N^{ab(12)}(p_2y^{ab(12)} + q_2y^{ba(12)}) + N^{b(12)}y^{b(12)} + N^{b(1)}y^{b(1)}.$$

(2)

De la même façon, pour le deuxième caractère, on a

$$Y^{(2)} = N^{a(2)}y^{a(2)} + N^{a(12)}y^{a(12)} + N^{ab(2)}(p_3y^{ab(2)} + q_3y^{ba(2)}) + N^{ab(12)}(p_4y^{ab(12)} + q_4y^{ba(12)}) + N^{b(12)}y^{b(12)} + N^{b(2)}y^{b(2)}$$

(3)

où

$$p_3 + q_3 = 1 \text{ et } p_4 + q_4 = 1.$$

2.1 Variance de l'estimateur

La variance conditionnelle des estimations $Y^{(1)}$ et $Y^{(2)}$ fondées sur la stratification a posteriori pour un sous-domaine donné qui ne tient pas compte de la correction d'échantillonnage pour population finie peut s'écrire de la façon suivante

$$V(Y^{(1)} | n^{a(1)}, n^{a(12)}, \text{ etc.}) = N_2^2 \frac{\sigma_2^2}{\sigma_{a(1)}^2} + N_2^2 \frac{n^{a(12)}}{\sigma_{a(12)}^2} + N_2^2 \frac{\sigma_2^2}{\sigma_{b(1)}^2} \left(p_1^2 \frac{n^{ab(1)}}{\sigma_{ab(1)}^2} + p_2^2 \frac{n^{ba(1)}}{\sigma_{ba(1)}^2} \right) + N_2^2 \frac{\sigma_2^2}{\sigma_{b(12)}^2} \left(p_2^2 \frac{n^{ab(12)}}{\sigma_{ab(12)}^2} + p_4^2 \frac{n^{ba(12)}}{\sigma_{ba(12)}^2} \right) + N_2^2 \frac{n^{b(12)}}{\sigma_{b(12)}^2}$$

(4)

La variance non conditionnelle de $Y^{(1)}$ est donnée approximativement par

$$V(Y^{(1)}) = \frac{n^A}{N^A} \left\{ N^{a(1)} \sigma_2^2 + N^{a(12)} \sigma_{a(12)}^2 + p_1^2 N^{ab(1)} \sigma_{ab(1)}^2 + p_2^2 N^{ab(12)} \sigma_{ab(12)}^2 + \frac{n^B}{N^B} \left\{ N^{b(1)} \sigma_2^2 + N^{b(12)} \sigma_{b(12)}^2 + p_2^2 N^{ba(1)} \sigma_{ba(1)}^2 + p_4^2 N^{ba(12)} \sigma_{ba(12)}^2 \right\} \right\}$$

(5)

qui est égale à la variance obtenue dans le cas d'un échantillonnage stratifié à répartition proportionnelle.

Les échantillons de taille n_a et n_b sont divisés en échantillons de taille n_a et n_b et de taille n_{ab} et n_{ba} de sorte que n_a et n_{ab} correspondent aux unités de l'échantillon de taille n_a appartenant aux domaines a et ab respectivement. De la même façon, les échantillons de taille n_b et n_{ba} sont les échantillons provenant de la division de l'échantillon de taille n_b en échantillons composés d'unités appartenant aux domaines b et ab respectivement. Dans une étude à plusieurs caractères, ces domaines seront une fois de plus divisés, générant des sous-domaines comme suit:

Soit $y^{(1)}$ et $y^{(2)}$ deux caractères devant faire l'objet d'une étude. Chacun des domaines habituels a , ab et b sera une fois de plus divisé en $a(1)$, $a(2)$, $ab(1)$, $ab(2)$ et $b(1)$, $b(2)$ respectivement. Dans ce cas-ci, $a(1)$, $a(2)$ et $ab(1)$ sont les sous-domaines composés des unités du domaine a ayant le caractère $y^{(1)}$, les deux caractères $y^{(1)}$ et $y^{(2)}$ et le caractère $y^{(2)}$ respectivement. La même explication s'applique aux autres sous-domaines $ab(1)$, $ab(2)$, etc. Par conséquent, la division des deux échantillons dans l'étude de deux caractères se fera de la façon suivante:

$$n_A = n_a + n_{ab}$$

où

$$n_a = n_{a(1)} + n_{a(2)} + n_{ab(1)} + n_{ab(2)}$$

et

$$n_B = n_b + n_{ba}$$

où

$$n_b = n_{b(1)} + n_{b(2)} + n_{ba(1)} + n_{ba(2)}$$

Dans ce cas-ci, les échantillons de taille $n_{a(1)}$, $n_{a(2)}$, etc. sont les échantillons composés des unités de l'échantillon de taille n_a appartenant respectivement aux sous-domaines $a(1)$, $a(2)$, etc. Si on se limite à un caractère, on peut alors définir

$$n_{A(1)} = n_{a(1)} + n_{ab(1)} + n_{ba(1)}$$

$$n_{B(1)} = n_{b(1)} + n_{ba(1)} + n_{ab(1)}$$

De la même façon, les échantillons de taille $n_{a(2)}$ et $n_{b(2)}$ sont définis pour le deuxième caractère. L'estimation du total pour le premier caractère est donnée par

$$Y^{(1)} = Y_{a(1)} + Y_{ab(1)} + Y_{ba(1)} + Y_{b(1)} + Y_{ab(2)} + Y_{ba(2)}$$

(1)

où $Y^{(1)}$, $Y^{(2)}$, etc. sont les totaux estimatifs pour le caractère $y^{(1)}$ des sous-domaines respectifs. Dans l'analyse qui suit, pour ce qui a trait aux domaines ayant les deux caractères, l'indice supérieur correspond au caractère examiné. Dans le cas des domaines ayant seulement un caractère, l'indice supérieur n'est pas utilisé parce que, de toute évidence, le domaine correspond au caractère.

Aussi, $p_1 + q_1 = 1$ et $p_2 + q_2 = 1$. Définissons $y^{a(1)}$, $y^{a(2)}$, etc. comme les moyennes des échantillons pour les sous-domaines respectifs ayant les caractères $y^{(1)}$ et $y^{(2)}$ respectivement.

Estimation du total pour deux caractères dans les enquêtes à bases de sondage multiples

B.C. SAXENA, P. NARAIN, et A.K. SRIVASTAVA¹

RÉSUMÉ

Le présent document traite de l'estimation de caractères multiples dans les enquêtes à bases de sondage multiples. Le gain attribuable à l'étude de deux caractères dans une même enquête par rapport à la tenue d'enquêtes distinctes pour chaque caractère est analysé. On fait également des comparaisons de coûts entre une enquête à bases multiples portant sur deux caractères et une enquête à base simple portant sur deux caractères.

MOTS CLÉS: Enquête portant sur plusieurs caractères; estimation par stratification a posteriori; optimisation; comparaison de coûts.

1. INTRODUCTION

La technique des enquêtes à bases multiples a été présentée pour la première fois par Hartley (1962) et examinée en détail par la suite par Lund (1968), Hartley (1974), Vogel (1975), Armstrong (1979), etc. Lund a proposé une solution de rechange à l'estimateur de Hartley, qui comporte la division de l'échantillon entre divers domaines. Hartley (1974) a poursuivi avec une approche plus générale pouvant s'appliquer à divers plans de sondage. Il a constaté que la plupart des cas possibles de bases multiples utilisaient des types d'unités différents dans leurs bases respectives. Bosecker et Ford (1976) ont développé l'estimateur de Hartley pour tirer avantage de la stratification dans le domaine de chevauchement. Serrurier et Phillips (1976) et Armstrong (1978) ont expérimenté les techniques comportant des bases de sondage multiples dans des enquêtes agricoles. L'utilité d'une enquête à bases de sondage multiples a été démontrée dans des cas très divers. Dans des enquêtes par sondage, il arrive parfois que l'intérêt ne réside pas seulement dans l'estimation d'un seul caractère, mais dans l'estimation de plusieurs caractères qui doivent être étudiés simultanément. Pour en arriver à une utilisation efficace des ressources, cela est souvent réalisé par la truchement d'enquêtes intégrées. Par exemple, dans le but d'estimer la production de récoltes de légumes, on peut concevoir une enquête unique qui a pour objet d'estimer la production de plusieurs récoltes de légumes. Également, au lieu d'utiliser une base composée de tous les producteurs de légumes, on peut utiliser une base incomplète mais relativement facilement accessible notamment une base composée des principaux producteurs de légumes. L'estimation du total pour deux caractères dans les enquêtes à bases multiples a été examinée dans le présent document. L'avantage d'étudier plusieurs caractères dans une seule enquête par rapport au recours à des enquêtes indépendantes portant chacune sur une caractéristique est également analysé.

2. ESTIMATEUR

Soit A et B deux bases de sondage empiétant l'une sur l'autre de taille N_A et N_B respectivement. Dans des enquêtes à bases multiples, deux échantillons de taille n_A et n_B sont tirés de façon indépendante par échantillonnage aléatoire simple à partir des bases de sondage A et B respectivement. Les bases chevauchantes génèrent des domaines a , b et ab définis comme suit:

- a : domaine composé des unités appartenant à la base A uniquement,
- b : domaine composé des unités appartenant à la base B uniquement,
- ab : domaine composé des unités appartenant à la fois à la base A et à la base B .

¹ B.C. Saxena, P. Narain, et A.K. Srivastava, Indian Agricultural Statistics Research Institute, Nouvelle-Delhi, Inde.

- McINNIS, R.M. (1977). Childbearing and land availability: some evidence from individual household data. *Population Patterns in the past*, (R.D. Lee ed.), New York: Academic Press, 201-227.
- MITCHELL, S.P., LING, D.G., et HANIS, E.H. (1982). *Final Report: Determination of Procedures and Costs for the Production of a Machine Readable Edition of the 1881 Census of Canada*. MAS contrat n° OSU80-00326.
- ORNSTEIN, M.D. (1978). The design of a sample of households from the 1871 census of Canada. *Manuscrit non publié*, Université York, Toronto.
- ORNSTEIN, M.D., et DARROCH, G.O. (1978). National mobility studies in past time: a sample strategy. *Historical Methods*, 11, 152-161.
- RAO, J.N.K., et SCOTT, A.J. (1981). The Analysis of categorical data from complex surveys: chi-square tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- SCOTT, A.J., et HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 17, 848-854.
- SMITH, D.S. (1978). A community-based sample of the older population from the 1880 and 1900 United States manuscript census. *Historical Methods*, 11, 67-74.
- SOLTOW, L. (1975). *Men and Wealth in the United States 1850-1870*. New Haven: Yale University Press.
- STATISTIQUE CANADA (1975). *Recensement du Canada de 1971: Bandes-échantillon à grande diffusion*. *Documentation des utilisateurs*, Ottawa.
- STATISTIQUE CANADA (1979). *Recensement du Canada de 1976: Bandes-échantillon à grande diffusion*. *Documentation des utilisateurs*, Ottawa.
- SWEIRENGA, R.P. (1983). Quantitative methods in rural landholding. *Journal of Interdisciplinary History*, 13, 787-808.

$n^{\text{ième}}$ élément est u , $v(u) = u$ pour $u = 1, \dots, N$. Le $u^{\text{ième}}$ élément désigne le $u^{\text{ième}}$ ménage d'une division. Supprimons maintenant tous les éléments de v qui correspondent aux ménages qui ont déjà été échantillonnés ou à ceux qui ne figurent pas sur le microfilm et réduisons le vecteur v à un vecteur w à $(N - M)$ dimensions. Les valeurs $w(u)$, $u = 1, \dots, N - M$ correspondent aux numéros de ménages susceptibles d'être échantillonnés. Dans l'algorithme, il suffit de redéfinir la taille de la population comme $N - M$ et la taille de l'échantillon comme $n - m$.

Avec de légères modifications, la méthode d'échantillonnage des ménages peut facilement produire un échantillon distinct et indépendant de personnes. Cette flexibilité repose sur le fait que chaque page des listes de recensement est numérotée et contient vingt-cinq noms. Le premier visionnement des microfilms doit servir à connaître le numéro de la dernière page de chaque division et le nombre de lignes que contient cette page. Avec l'algorithme de Bebbington, l'ordinateur imprimera le numéro de la page et le numéro de la ligne où figure le nom de la personne échantillonnée.

Cette méthode d'échantillonnage a été programmée et testée avec succès par le Centre de données sur les sciences sociales. Selon l'étude de faisabilité, par exemple, la recherche des unités échantillonnées a représenté environ 6% de la durée totale estimée de la saisie de données pour l'échantillon de ménages et 18,5% de cette durée pour l'échantillon de personnes. Voir Mitchell *et coll.* (1982, p. 20-21).

BIBLIOGRAPHIE

BATEMAN, F., et FOUST, J.D. (1974). A sample of rural households selected from the 1860 manuscript censuses. *Agricultural History*, 48, 75-93.

BEBBINGTON, A.D. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 135.

BODE, F.A., et GINTER, D.E. (1984). A critique of land holding variables in the 1860 census and the Parker-Gallman sample. *Journal of Interdisciplinary History*, 15, 277-295.

DARROCH, A.G., et ORNSTEIN, M.D. (1980). Ethnicity and occupational structure in Canada in 1871: the vertical mosaic in historical perspective. *Canadian Historical Review*, 61, 305-333.

FOGEL, R.W., et ENGERMAN, S.L. (1974). *Time on the Cross: Evidence and Methods*. Boston: Little, Brown and Co.

FOUST, J.D. (1975). *The Yeoman Farmer and Westward Expansion of U.S. Cotton Production*. New York: Arno Press.

GRAHAM, S.N. (1980). *1900 Public Use Sample: User's Handbook*. Seattle: Centre for Studies in Demography and Ecology, University of Washington.

HAMMARBERG, M.A. (1971). Designing a sample from incomplete historical lists. *American Quarterly*, 23, 542-561.

HAMMARBERG, M.A. (1977). A sampling design for Mormon Utah, 1880. *Journal of Interdisciplinary History*, 7, 453-476.

HOLT, D., SCOTT, A.J., et EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, série A*, 143, 303-320.

JOHNSON, R.C. (1978a). A procedure for sampling manuscript census schedules. *Journal of Interdisciplinary History*, 8, 513-530.

JOHNSON, R.C. (1978b). The 1900 census sampling project: methods and procedures for sampling and data entry. *Historical Methods*, 11, 147-151.

Le numéro du premier ménage à échantillonner apparaît à l'écran. Le codeur fait alors avancer le film jusqu'au numéro de ménage voulu. Une fois les données de ce ménage saisies, il appuie sur une touche appropriée et le deuxième numéro de ménage apparaît à l'écran. Lorsqu'il a fini d'échantillonner cette division, il appuie sur la même touche pour passer à une autre division. Il peut parfois y avoir des ménages manquants. Il se peut, par exemple, qu'une feuille d'agent recenseur contenant vingt-cinq noms ait été perdue. Quand un codeur constate l'absence d'un ménage en échantillonnant une division, il introduit en mémoire le numéro correspondant en indiquant qu'il s'agit d'un ménage manquant et fait de même avec tous les autres numéros sous lesquels il ne voit aucun ménage d'inscrit. Il poursuit ensuite son échantillonnage jusqu'à ce qu'il ait couvert toute la division. Comme il manque au moins un ménage dans l'échantillon prévu, le codeur doit rembobiner le microfilm et poursuivre l'échantillonnage dans la même division. Le principal avantage de cette méthode est de permettre au codeur de faire défiler les microfilms dans une seule direction, sauf lorsqu'il y a des ménages manquants.

L'algorithme utilisé dans cette méthode d'échantillonnage est fondé sur un fichier contenant des données sur les divisions ou groupes de divisions de même que sur l'algorithme de Bebbington (1975) conçu pour le prélèvement d'un échantillon aléatoire simple sans remise. Après le premier visionnement des microfilms, on crée un fichier indiquant l'identificateur de la division et le nombre de ménages que contient la division. Si les divisions sont groupées, leur taille respective est enregistrée. Il suffit pour le codeur de taper l'identificateur de la division à échantillonner pour connaître la taille de cette division. Pour obtenir une répartition proportionnelle de l'échantillon de ménages, il faut que la taille de l'échantillon soit égale au produit de la taille de la division par la fraction de sondage s'appliquant à l'ensemble du recensement. Avec l'algorithme de Bebbington (1975), on effectue ensuite sondage progressif des unités de l'échantillon à partir d'une liste contenant dans l'ordre les numéros des ménages de la division donnée. Les numéros de ménage sont testés tour à tour, puis choisis ou rejetés. Lorsqu'un numéro de ménage est choisi, il est affiché sur l'écran et la sélection s'interrompt le temps de la saisie des données. Comme les numéros sont choisis dans l'ordre ascendant, le codeur n'a qu'à faire défiler le film dans une seule direction pour trouver les documents manuscrits voulus.

Cet algorithme permet aussi d'échantillonner des strates combinées ou divisions groupées. Supposons que L strates de taille N_1, \dots, N_L aient été combinées dans une seule strate de taille $N = N_1 + N_2 + \dots + N_L$. Il suffit d'utiliser les tailles des strates pour connaître le ménage échantillonné dans chaque strate. Supposons, dans l'algorithme, que les unités $s(1), \dots, s(n)$ aient été choisies pour former l'échantillon $1 \leq s(i) \leq N$. Si pour tout i ($i = 1, \dots, n$) $N_1 + N_{h-1} + \dots + N_{h-1} < s(i) \leq N_1 + N_{h-1} + \dots + N_{h-1} + N_h$, où $N_1 + N_{h-1} + \dots + N_{h-1} = 0$ pour $h = 1$, l'unité $s(i)$ se trouve dans la strate h et le numéro de ménage correspondant est $s(i) - (N_1 + \dots + N_{h-1})$.

On peut aussi modifier l'algorithme de sorte qu'il tienne compte des ménages manquants. La méthode décrite ci-dessous ne suppose pas l'énumération des ménages manquants avant l'échantillonnage, une fois terminée l'échantillonnage d'une strate au moyen de l'algorithme de Bebbington (1975), deux cas sont possibles: ou bien on a constaté l'absence d'un certain nombre de ménages, ou bien on n'a constaté aucune absence. Le deuxième cas ne pose aucun problème, l'échantillonnage étant complet. Dans le premier cas, la taille de l'échantillon formé, disons m , est inférieure à la taille voulue n . Pour obtenir un échantillon de ménages existants de taille n il faut échantillonner $n - m$ ménages additionnels. À cette fin, nous recommandons l'échantillonnage de la strate, mais cette fois avec une liste des ménages qui ont déjà été échantillonnés et des ménages manquants connus. Supposons que la liste contienne M ménages ($M \geq n$: le codeur peut avoir remarqué et enregistré des ménages manquants qui n'ont pas été échantillonnés). Définissons un vecteur v à N dimensions, où la valeur du

Canada ont voulu intégrer toutes les données du recensement dans une base ordiologique ou créer des bandes-échantillons à grande diffusion semblables à celles qui avaient été produites pour les recensements de 1971 et de 1976 (voir Statistique Canada (1975, 1979) pour la documentation). Le Centre de données sur les sciences sociales de l'Université Western Ontario s'est donc vu accorder un contrat pour réaliser une étude de faisabilité, et on a demandé à l'auteur d'élaborer une méthode d'échantillonnage qui permettrait de former l'échantillon à grande diffusion. La présente section contient une description du plan de sondage proposé. On trouvera un rapport de l'étude de faisabilité dans Mitchell *et coll.* (1982). Le questionnaire I contient des renseignements sur l'âge, le sexe, le pays de naissance, l'origine ethnique, la profession et l'état matrimonial de chaque personne et indique si cette personne souffre d'incapacité. Les sept autres questionnaires contiennent des renseignements sur l'industrie, l'agriculture, les forêts, la pêche et les mines. On trouvera une brève description des questionnaires dans *Recensement du Canada 1880-1881*, volume I, p. v-xv.

Nous décrivons brièvement ci-dessous les conditions de base des échantillons à grande diffusion des questionnaires dans *Recensement du Canada 1880-1881*, volume I, p. v-xv. Nous décrivons ces échantillons soient conformes à ceux de 1971 et de 1976, il faudrait avoir deux échantillons indépendants, soit un échantillon de ménages et un échantillon de personnes. Si toutefois il n'était économiquement possible de produire qu'un échantillon, ce devrait être en priorité un échantillon de ménages. L'échantillon à grande diffusion tiré du recensement des Etats-Unis de 1900 et décrit par Johnson (1978b) et Graham (1980) est un échantillon de ménages. En outre, il semble que les historiens préfèrent avant tout le ménage comme unité d'échantillonnage. L'échantillon du recensement de 1900 indique également qu'une taille d'échantillon de l'ordre de 100,000 unités serait souhaitable pour l'échantillon de personnes ou l'échantillon de ménages. En ce qui concerne le recensement du Canada de 1881, cela impliquerait une fraction de sondage d'environ 2% pour l'un ou l'autre échantillon. Enfin, il est également souhaitable d'utiliser, pour l'un et l'autre échantillon, un échantillonage stratifié avec répartition proportionnelle où les strates correspondent à des régions géographiques. Cette méthode est celle qui a été le plus couramment utilisée jusqu'à maintenant par les historiens, et elle produit un échantillon autopondéré. Le choix des unités dans une strate devrait se faire par échantillonnage aléatoire simple plutôt que par échantillonnage systématique. Johnson (1978a) soutient que l'échantillonnage systématique, bien que pratique, ne convient pas aux questionnaires manuscrits de recensement. Des voisins ont des caractéristiques similaires et ne pourraient jamais appartenir à un même échantillon systématique. Or les historiens pourraient vouloir étudier les personnes qui ont des caractéristiques similaires.

Compte tenu de ces conditions de base, nous proposons pour le prélèvement de l'échantillon de ménages un échantillonnage aléatoire stratifié où les strates correspondent, comme dans Ornstein (1978), aux divisions de recensement (secteurs de dénombrement actuels) plutôt qu'aux bobines de microfilm utilisées par Johnson (1978b) et Graham (1978). Les divisions de recensement sont des strates géographiques naturelles. En outre, les ménages sont numérotés successivement sur les listes des agents recenseurs, chaque page manuscrite comportant vingt-cinq noms. Il suffirait alors de faire un premier visionnement des microfilms pour connaître le nombre de ménages dans chaque strate. Avec une fraction de sondage de 2 à 2,5% et une répartition proportionnelle, on obtient des échantillons de moins de deux ménages dans des divisions (strates) comptant un peu moins de cent ménages. Dans ce cas, la division en question devrait être intégrée à des divisions contiguës. Une stratification poussée au-delà de la division, comme celle d'Ornstein (1978), paraît inutile et ferait monter sensiblement le coût de l'échantillonnage.

L'échantillonnage peut être facilement informatisé. Pour un codeur assis à un terminal et utilisant une visionneuse de microfilms, l'échantillonnage est une opération simple. Lorsqu'un codeur échantillonne une division, il n'a qu'à taper le code de la division voulue et

le nombre d'enfants par famille et l'abondance des terres dans certaines régions. Il a commencé par répartir environ 300 cantons en strates selon l'année de colonisation. Il a ensuite prélevé un échantillon de cantons dans les strates et des échantillons de fermes dans les cantons. Il semble que McGinnis ait choisi l'échantillonnage à deux degrés pour des raisons d'économie. En effet, comme les fermes échantillonnées étaient ensuite apparées au dossier correspondant dans le recensement de l'agriculture, il était moins long et, par conséquent, moins coûteux d'échantillonner quelques cantons et d'apparier les dossiers de plusieurs fermes d'un canton que de stratifier les cantons et d'apparier les dossiers d'un petit nombre de fermes dans chaque strate. Le même raisonnement s'applique aux travaux d'Hammarberg (1977). Lui aussi rattachait d'autres dossiers au ménage échantillonné.

2.5 Échantillonnage en grappes à deux degrés stratifié

Smith (1978) a appliqué une méthode d'échantillonnage en grappes à deux degrés stratifié pour étudier la population âgée dans le recensement des États-Unis de 1900. Les strates sont définies comme les régions de recensement, les comtés qui forment les régions étant les unités primaires d'échantillonnage. Celles-ci sont choisies selon une probabilité proportionnelle à la taille de leur population. Pour chaque comté, Smith a prélevé plusieurs pages de questionnaires du recensement. Il a ensuite relevé les noms des personnes de plus de 50 ans qui figuraient sur chacune des pages échantillonnées. L'échantillonnage en grappes était nécessaire du fait qu'il aurait été trop coûteux de déterminer toutes les personnes qui pouvaient être échantillonnées. Smith tente par la même occasion de comparer quelques distributions d'échantillons aux données publiées du recensement. Il se sert pour cela de la fonction des observations normalement utilisée dans les tests d'hypothèses portant sur une proportion simple même s'il s'agit de données multinomiales.

Foust (1968, chap. 2) décrit un deuxième cas d'échantillonnage en grappes à deux degrés stratifié. L'échantillon, dit échantillon Parker-Gallman, a été tiré du recensement des États-Unis de 1860 pour étudier les régions productrices de coton du sud du pays. Les strates étaient 405 *régions cotonnières* du Sud, où l'on avait produit au moins 1,000 balles de coton de 400 livres dans les douze mois qui précédaient le jour du recensement. Pour chaque région, on a prélevé un échantillon aléatoire systématique de pages de documents manuscrits du recensement, et un groupe de cinq plantations a été choisi au hasard sur une page donnée, le groupe étant considéré comme une grappe. On a recouru à l'échantillonnage en grappes car il fallait consulter trois questionnaires de recensement différents pour recueillir des données sur une plantation particulière. L'appariement des questionnaires a été qualifié de très laborieux. Fogel et Engerman (1974, p. 22-25) énumèrent plusieurs autres échantillons se rattachant à l'échantillon Parker-Gallman. Bode et Ginter (1984) forment des critiques sur le contenu de l'échantillon.

De tous les échantillons considérés ici, l'échantillon Parker-Gallman et les échantillons de Bateman et Foust (1974) sont ceux qui ont été le plus étudiés. Swierenga (1983) a passé en revue une bonne partie des travaux fondés sur ces échantillons.

3. ÉCHANTILLONS À GRANDE DIFFUSION TIRÉS DU RECENSEMENT DU CANADA DE 1881

Au début des années 1980, les Archives publiques du Canada ont obtenu les copies du *questionnaire 1: Renseignements d'ordre général* du recensement du Canada de 1881. Les questionnaires ont été microfilmés, et on peut maintenant trouver des exemples de ces microfilms dans la plupart des bibliothèques de collège ou d'université et dans beaucoup de bibliothèques publiques. Après avoir produit les microfilms, les Archives publiques du

qu'un seul visionnement. Elle élimine en outre le problème des strates vides ou des strates à une unité lorsque la fraction de sondage pour une strate est faible. En revanche, comme elle ne comporte qu'un seul visionnement, il faut pouvoir résoudre de façon ponctuelle les principaux problèmes qui peuvent survenir.

2.3 Échantillonnage en grappes stratifié

Bateman et Foust (1974) ont obtenu un échantillon des exploitations agricoles du nord des États-Unis à l'aide des données du recensement de 1860. Ils ont divisé le Nord en deux strates, est et ouest, et prélevé un échantillon aléatoire de comtés ruraux dans chaque strate. Ensuite, ils ont choisi au hasard, dans chaque comté, une commune rurale (grappe) et recueilli des données sur toutes les exploitations agricoles situées sur le territoire de la commune. Une des raisons du choix de l'échantillonnage en grappes est que cette méthode est économique. Comme les données sur les exploitations agricoles provenaient du recensement de l'agriculture et que les données démographiques sur les propriétaires de ces exploitations agricoles et ceux qui y travaillaient provenaient du recensement de la population, il était plus facile de faire le lien entre les exploitations agricoles et leurs propriétaires respectifs en se limitant à une commune. Swierenga (1983) donne une deuxième raison pour justifier l'utilisation de l'échantillonnage en grappes. Il affirme que les données recueillies pour une commune ont permis d'estimer la productivité globale des facteurs en agriculture et de définir toute la main-d'œuvre agricole, y compris les travailleurs agricoles qui demeurent à l'extérieur des 12,000 exploitations comprises dans l'échantillon (p. 793). Comme les grappes (communes) n'ont pas été choisies selon une probabilité proportionnelle à la taille, le plan de sondage n'a pas produit d'échantillons autopondérés.

Bateman and Foust (1974) ont aussi appliqué quelques tests pour vérifier la représentativité de leur échantillon. À l'instar d'Hammarberg (1971), ils ont appliqué le test du khi carré pour comparer les observations de l'échantillon aux observations probables de la population. Dans le cas des variables continues, ils ont utilisé le test t. Les estimations de la moyenne et de la variance étaient des estimations *simples*, c'est-à-dire qu'elles n'étaient pas fondées sur le plan de sondage.

2.4 Échantillonnage à deux degrés stratifié

Hammarberg (1977) a appliqué une méthode d'échantillonnage à deux degrés stratifié pour échantillonner des ménages dans le recensement du territoire de l'Utah de 1880. Les strates sont un amalgame assez complexe de cinq régions géographiques de l'Utah, de quelques comtés de régions populaires et de quelques grandes municipalités. Hammarberg a prélevé dans chaque strate un échantillon de municipalités ou de wards. Les municipalités qui constituaient déjà des strates étaient automatiquement incluses dans l'échantillon. Les wards sont à l'Eglise mormone ce que les paroisses étaient à l'Eglise chrétienne médévale. Un échantillon de ménages a ensuite été prélevé dans les municipalités ou les wards choisis. Cet échantillon était autopondéré. L'argument que fait valoir Hammarberg à la page 460 de son ouvrage pour justifier ce mode de stratification est convaincant:

«Comme le mode d'organisation fondamental de la population repose sur une structure géographique et que la plupart des documents officiels – civils et religieux – correspondent à cette structure, on peut dire qu'un échantillonnage de la population suivant une répartition géographique équivalait dans une large mesure à un échantillonnage des dossiers produits pour cette population.»

McInnis (1977) a aussi eu recours à l'échantillonnage à deux degrés stratifié pour obtenir un échantillon des dossiers du recensement colonial de 1861. Il étudiait alors la relation entre

Soltow (1975) s'est servi d'échantillons des recensements américains de 1850, de 1860 et de 1870 pour faire une étude de la richesse aux États-Unis. Pour chaque année de recensement, il a prélevé un échantillon sur chaque bobine de microfilm de telle sorte que l'échantillon soit stratifié par bobine, ce qui équivaut à peu près à une stratification géographique. Le plan de sondage de Soltow semble correspondre à un échantillonnage systématique. Pour prélever un échantillon, il a déterminé un point sur l'écran de la visionneuse de microfilms puis a visionné le film. Il faisait avancer la pellicule par demi-tours successifs de manière jusqu'à ce qu'un questionnaire de recensement soit relativement centré sur le point désigné sur l'écran. Pour être sélectionné, le questionnaire devait concerner une personne de sexe masculin âgée de 20 ans ou plus. En outre, l'échantillon du recensement de 1860 comprenait 40 fois plus de personnes dont l'âge était inférieur à \$100,000 ou plus que de personnes dont l'âge était inférieur à \$100,000 (p. 5), de sorte que cet échantillon n'était pas autopondéré. Bien que Soltow n'en fasse pas mention, il peut avoir voulu *suréchantillonner* les personnes plus fortunées de manière à disposer d'un échantillon suffisamment grand pour lui permettre d'établir des comparaisons avec les classes moins aisées de la société. Il a également comparé les observations de ses échantillons aux distributions de fréquence publiées mais n'a effectué aucun test de validité de l'ajustement. Il a constaté que pour diverses variables, les données des échantillons se rapprochaient sensiblement de celles des recensements sur le plan des moyennes et des proportions. Ces observations s'appliquaient même à des variables comme le patrimoine moyen, chose surprenante compte tenu du *suréchantillonnage* de personnes plus fortunées et du fait que l'estimation de Soltow semble correspondre à la moyenne de l'échantillon.

Dartoch et Ornstein (1980) ont utilisé un échantillon du recensement du Canada de 1871 pour étudier la relation entre l'origine ethnique et la profession. La méthode d'échantillonnage utilisée est décrite dans Ornstein (1978). Pour les besoins des deux études, il a fallu *suréchantillonner* certains groupes ethniques de manière à ne pas obtenir un échantillon autopondéré. Ne tenant pas compte de ce *suréchantillonnage*, les deux historiens ont appliqué un échantillonnage aléatoire stratifié, la stratification étant fondée sur la structure géographique-hiérarchique des dossiers du recensement: provinces, districts, sous-districts et divisions. La division correspond au secteur de dénombrement actuel et semble être le critère naturel de stratification. Toutefois, Ornstein (1978) l'a subdivisée en strates et a prélevé deux ménages dans chaque strate. Il ne dit pas comment il opère cette subdivision mais il la justifie en affirmant que l'échantillonnage de deux unités par strate réduit au minimum la variance des estimations de certaines valeurs d'une population. Bien qu'Ornstein (1978) ne le précise pas, il semble qu'il ait voulu accroître l'efficacité de la stratification en formant des strates à l'intérieur d'une division aussi homogène que possible. Il s'est ainsi trouvé à accroître le coût de l'échantillonnage. Enfin, sa méthode exigeait que l'on fasse défiler le film au moins deux fois dans la visionneuse, la première fois pour obtenir le nombre de ménages par division et la deuxième pour échantillonner les ménages.

Dans leurs ouvrages, Johnson (1978b) et Graham (1980) indiquent comment ils ont obtenu un échantillon à grande diffusion du recensement des États-Unis de 1900. Dans un autre ouvrage, Johnson (1978a) reprend à peu près les mêmes éléments pour décrire comment il a échantillonné les questionnaires du recensement du Rhode Island de 1860. Dans les trois cas, l'échantillon est formé en déterminant au hasard des lignes sur le microfilm et en cherchant par la suite ces lignes à l'aide d'une visionneuse munie d'un compteur. Compte tenu de la méthode d'échantillonnage, la taille globale de l'échantillon est aléatoire. Graham (1980, p. 41) donne un certain nombre de critères permettant d'accepter ou de rejeter les lignes échantillonnées. Il s'agit d'un échantillonnage aléatoire stratifié et les bobines de microfilms servent de strates. La stratification est fondée sur une répartition géographique dans la mesure où les questionnaires de recensement d'une même région se trouvent tous sur la même bobine de microfilms. Cette méthode a l'avantage de rendre l'exécution efficace et de ne nécessiter

2. REVUE RÉTROSPECTIVE

Dans la troisième partie de l'exposé, nous proposons une méthode permettant de prélever des questionnaires du recensement du Canada de 1881 dans le but de créer une bande-échantillon à grande diffusion. Cette recherche faisait partie d'un projet exécuté pour les Archives publiques du Canada. Elle a été confiée par contrat au Centre de données sur les sciences sociales de l'Université Western Ontario. Nous donnons ici une description du plan d'échantillonnage. On trouvera un rapport complet sur le projet dans Mitchell *et coll.* (1982). Le plan d'échantillonnage utilisé ressemble à certains égards aux plans qui ont servi à la création des bandes-échantillons à grande diffusion pour les recensements de 1971 et de 1976. Les plans d'échantillonnage reposent tous sur la stratification; pour le recensement de 1881, toutefois, la stratification n'a pu se faire qu'en fonction d'une répartition géographique. Les travaux portant sur l'échantillonnage des documents manuscrits de recensement peuvent être classés suivant la méthode d'échantillonnage utilisée. Nous décrivons ci-dessous les diverses méthodes utilisées par ordre croissant de complexité du plan de sondage.

2.1 Echantillonnage en grappes

Ornstein et Darroch (1978) ont proposé une méthode simple et économique permettant d'échantillonner des dossiers de recensement et de les raccorder d'une période à une autre. Cette méthode consiste essentiellement à former des grappes de noms de famille et à constituer des échantillons à partir de ces grappes. Les grappes sont désignées par la première lettre du nom de famille. Si les mêmes grappes sont échantillonnées dans divers recensements à la fois, une personne dont le nom figure dans plus d'un recensement fera partie de l'échantillon choisi. Il y a donc moins de cas à analyser pour la liaison et le coût est par conséquent moins élevé. Ce plan de sondage se prête particulièrement bien aux études chronologiques de la migration ou de l'évolution démographique.

2.2 Echantillonnage stratifié

Aucun des plans de sondage avec stratification considérés dans la présente section ne prévoit une répartition optimale de l'échantillon. Cette situation s'explique par le fait qu'aucun des historiens n'était en mesure de connaître a priori les variations à l'intérieur des strates. Pour obtenir ce genre de renseignements, il aurait fallu supporter une hausse sensible du coût de chaque projet.

Hammarberg (1971) a utilisé une méthode de sondage à deux phases, ou méthode d'échantillonnage double, dans l'espoir de réduire le biais engendré par l'échantillonnage d'un ensemble incomplet de dossiers. Les dossiers échantillonnés dans la deuxième phase ont été les autres commerciaux de neuf comtés de l'Indiana. Dans la première phase du sondage, Hammarberg a prélevé un échantillon dans un ensemble de dossiers supposé complet, le recensement des États-Unis de 1870. Il a appliqué la méthode d'échantillonnage aléatoire stratifié avec répartition proportionnelle de sorte que l'échantillon soit autopondéré. Les strates correspondaient aux neuf comtés de l'Indiana. On trouve deux aspects de cette recherche dans des études ultérieures sur l'échantillonnage de documents anciens. Les strates correspondent à des régions géographiques et l'échantillon est autopondéré.

Hammarberg (1971) a également appliqué le test du khi carré classique à certaines variables pour vérifier dans quelle mesure la distribution des données de l'échantillon se rapprochait des distributions de population établies à partir des données du recensement. Dans beaucoup d'autres études, on ne s'est pas préoccupé de vérifier la représentativité de l'échantillon.

Echantillonnage des questionnaires manuscrits de recensement reproduits sur microfilm

D.R. BELLHOUSE¹

RÉSUMÉ

Dans la première partie du document, nous procédons à une revue rétrospective des travaux d'historiens sur les dossiers manuscrits de recensement microfilmés. Les historiens ont utilisé plusieurs types de plan d'échantillonnage qui vont, par ordre croissant de complexité, de l'échantillonnage aléatoire en grappes et stratifié jusqu'à l'échantillonnage en grappes à deux degrés stratifié. Dans la deuxième partie, nous proposons une méthode permettant de créer une bande-échantillon à grande diffusion qui contiendrait des données du recensement du Canada de 1881. Cette recherche faisait partie d'un projet pilote exécuté pour les Archives publiques du Canada et a été réalisée par le Centre de données sur les sciences sociales de l'Université Western Ontario. Le projet pilote avait pour but de déterminer s'il était avantageux et possible, du point de vue économique et technique, de construire une base de données ordonnée à l'aide des questionnaires microfilmés du recensement du Canada de 1881.

MOTS CLÉS: Échantillonnage aléatoire informatisé; dossiers microfilmés; sondages à plusieurs degrés; échantillons à grande diffusion; stratification.

1. INTRODUCTION

Pour écrire l'histoire d'une personne ou d'un peuple, l'historien a besoin de sources. De nos jours, beaucoup d'historiens veulent retracer la vie de *l'homme de la rue*. Pour mener à bien leur recherche, ils peuvent se servir de documents comme les questionnaires de recensement, les titres de propriété et les annuaires commerciaux. Dans le présent document, nous nous penchons sur l'utilisation des questionnaires de recensement comme source. Le principal inconvénient que présente l'utilisation de données de recensement est l'abondance de ces données. L'historien qui dispose d'un budget de recherche normal n'a ni le temps ni les ressources financières ou humaines nécessaires pour passer en revue tous les questionnaires du recensement. Pour contourner cette difficulté, il doit former un échantillon aléatoire de questionnaires. La plupart des questionnaires de recensement que peut consulter l'historien sont reproduits sur microfilm. Au Canada, les questionnaires reproduits sur microfilm sont ceux des recensements coloniaux de 1841, de 1851 et de 1861 et ceux des recensements du Canada de 1871 et de 1881. Le problème se résume donc pour l'historien à déterminer le plan de sondage approprié pour prélever un échantillon de questionnaires microfilmés. Dans la deuxième section du document, nous passons en revue les méthodes d'échantillonnage appliquées par les historiens. L'application de ces méthodes a donné des résultats très inégaux. Dans certains cas, les résultats ont été très satisfaisants, les historiens ayant su adapter l'application des méthodes à l'objet de la recherche. Dans d'autres cas, toutefois, il semble que les historiens aient utilisé des plans de sondage inutilement complexes. Un plan de sondage complexe peut avoir des effets qui s'écartent sensiblement de l et, par conséquent, compliquer l'analyse des données. Voir, par exemple, Rao et Scott (1981) et Holt et coll. (1980) pour des commentaires sur l'analyse de données qualitatives, et Scott et Holt (1982) pour des commentaires sur l'analyse par régression. Enfin, les rapports de plusieurs des enquêtes analysées ci-dessous ne contiennent pas assez de renseignements pour nous permettre d'évaluer les raisons qui ont justifié le choix d'un plan de sondage particulier.

¹ D.R. Bellhouse, Département des sciences statistiques et actuariales, Université Western Ontario, London Ontario, Canada N6A 5B9.

BIBLIOGRAPHIE

- CHOUDHRY, G.H., LEE, H., et DREW, J.D. (1985). Optimisation du coût et de la variance dans le cadre de l'enquête sur la population active du Canada. *Techniques d'enquête*, 11, 37-56.
- DAHMSSTRÖM, P., et HAGNELL, M. (1978) The formation of strata using cluster analysis. Document interne, Department of Statistics, University of Lund, Sweden.
- FOY, P. (1984). Programme de stratification pour l'enquête sur la population active du Canada: Guide de l'utilisateur. Document interne, division des méthodes de recensement et d'enquêtes ménages, Statistique Canada.
- FRIEDMAN, H.P., et RUBIN, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.
- JUDKINS, D.R., et SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 274-284.
- KOSTANICH, D., JUDKINS, D.R., SINGH, R.P., et SCHANTZ, M. (1981). Modification of Friedman-Rubin's clustering algorithm for use in stratified PPS sampling. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 285-290.
- MAYDA, F., DREW, J.D., et LINDEYER, J. (1985). Phase-in of the redesignated Labour Force Survey. Document interne, division des méthodes de recensement et d'enquêtes ménages, Statistique Canada.
- PLATEK, R., et SINGH, M.P. (1976). Méthodologie de l'enquête sur la population active du Canada. No. 71-526 au catalogue, Statistique Canada.
- SINGH, M.P., DREW, J.D., et CHOUDHRY, G.H. (1984). Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981. *Techniques d'enquête*, 10, 139-154.

(iii) la stratification à deux niveaux mène à une meilleure représentation de la variance de réponse corrélative dans les estimations de la variance. Dans l'ancien plan de sondage, il n'y avait habituellement qu'un seul intervieweur par strate, ce qui occasionnait une sous-estimation de cette composante de la variance. Avec des strates secondaires non-géographiques, mais des tâches d'intervieweurs toujours géographiques, ce problème sera moins fréquent.

Les contraintes de coûts associées à l'exécution des programmes informatiques impliqués nous ont obligé à traiter certaines U.A.R. de façon particulière. En effet, la région de Montréal se voit diviser en sept parties indépendantes, le temps de la stratification. Il en est de même avec Toronto (5 parties), Winnipeg (2 parties), Calgary (2 parties), Edmonton (2 parties) et Vancouver (3 parties). Ces divisions se sont effectuées à partir de critères "naturels" tels que suggérés par la géographie de ces régions.

Dans les grandes U.A.R., les immeubles d'appartements existant au moment de la conception du plan de sondage ont été triés en fonction de strates primaires dans lesquelles ils étaient situés, pour atteindre une stratification implicite de cet échantillon.

Dans les U.A.R. de taille moyenne, où l'échantillon était insuffisant pour justifier la stratification à deux niveaux, on s'est contenté de construire des strates optimales à l'aide de l'algorithme de stratification, sans aucune contrainte géographique.

Les plus petites U.A.R., celle qui n'ont pas été décomposées en côtes d'ilot aux fins du recensement, ont été stratifiées manuellement, sans souci d'optimalité.

Mentionnons finalement que la période d'introduction du nouvel échantillon nous a amené une contrainte supplémentaire. Pour les grandes U.A.R., on a défini le secteur noyau comme étant constitué des strates complètes de l'ancienne conception qui n'ont pas été affectées par des changements de frontières. En s'assurant que les frontières des strates de la nouvelle conception respectent ces secteurs noyau, on s'est aussi assuré que le nouvel échantillon représente la même aire géographique que l'ancien échantillon durant la période d'introduction. Ceci a permis le remplacement graduel de l'ancien échantillon par le nouveau et par le fait même évité un coûteux chevauchement du nouvel échantillon et de l'ancien échantillon (Mayda, Drew, et Lindey 1985).

5. CONCLUSION

L'utilisation de l'algorithme de classification multi-variée nous a permis de développer une stratification très générale, renforçant ainsi l'E.P.A. dans son rôle d'enquête-ménages générale. De plus, l'automatisation des étapes de stratification dans les parties N.A.R. et A.R., et de la délimitation des U.P.E. dans les U.N.A.R., a mené à une réduction significative dans le coût et le temps nécessaire pour remanier l'échantillon.

Le système est documenté (Foy 1984) et il peut être utilisé pour la formation de strates ou de grappes dans les autres enquêtes. On pourrait également s'en servir dans des situations où on a à définir des régions statistiques ou administratives, en utilisant toute une gamme de variables. Pour l'E.P.A., un aspect de recherche à approfondir serait le choix entre des strates contiguës et non-contiguës, et les implications des strates non-contiguës sur le plan d'échantillonnage.

REMERCIEMENTS

Les auteurs désirent remercier Sylvie Trudel et Marc Joncas pour leur aide apportée à la réalisation des études mentionnées dans ce rapport, ainsi que les membres du Comité du remaniement de l'E.P.A. pour leurs précieuses suggestions. Ils remercient aussi l'arbitre pour ses commentaires utiles.

Tableau 5
Comparaison de trois méthodes de stratification (U.A.R.)

Variables	Ancienne conception		Stratification à deux niveaux (option 1)		Stratification compacte (option 2)	
	1971	1981	1971	1981	1971	1981
Agriculture	5.5	2.9	3.2	1.8	3.4	1.8
Foresterie	2.2	2.3	2.1	1.7	2.2	2.3
Mines	7.6	4.9	8.5	4.1	7.6	4.0
Manufacture	34.7	35.0	36.6	34.1	39.1	35.0
Construction	32.5	29.6	39.7	30.1	42.4	33.4
Transport	9.2	6.8	18.1	11.6	20.0	11.6
Services	29.5	27.5	45.8	33.1	46.7	32.1
Employés	15.1	8.0	31.4	14.1	32.8	12.6
Chômeurs ^a	14.6	5.7	14.9	6.7	15.5	7.1
Revenu	39.4	38.6	51.8	29.8	53.6	48.0
Population 15-24	9.6	15.2	12.5	17.5	13.3	14.9
Population 55 +	27.9	18.3	34.0	20.8	32.6	18.5
Ménages 1 personne	20.3	19.2	36.3	33.8	37.8	35.0
Ménages 2 personnes	21.9	20.3	40.3	30.9	40.1	30.2
Logements possédés	20.3	15.3	29.7	22.9	32.1	24.9
Education secondaire	32.6	42.4	50.3	47.9	51.6	49.1
Population 15 + ^a	27.0	8.2	38.0	13.4	37.6	12.0
Logements ^a	21.8	18.5	41.7	33.8	42.1	34.3

^a Non utilisé comme variable de stratification.

On remarque également que les trois méthodes donnent une stratification généralement robuste au fil du temps, telle que reflétée par la comparaison entre les indices 1981 et 1971. Des exceptions importantes à cette règle, malheureusement, semblent être les caractéristiques employées et chômeurs.

4.3 Nouvelle conception

Étant donnée la similitude des résultats entre les deux options étudiées, on a décidé d'adopter la stratification à deux niveaux (option 1) dans les grandes villes où l'échantillon est de 300 ménages ou plus, pour les raisons suivantes:

- (i) le fait d'avoir la contiguïté au niveau des strates primaires nous donne une unité convenable pour la mise à jour de l'échantillon

- (ii) les strates primaires pourront être utilisées pour la formation des tâches d'intervieweur. La taille des strates a été déterminée de manière à ce que l'échantillon à l'intérieur du secteur, soit l'échantillon aréolaire plus l'échantillon provenant des strates d'appartements, corresponde à deux tâches d'intervieweur (160 ménages dans le coeur de la ville et 120 ailleurs).

4. STRATIFICATION DANS LES UNITÉS AUTOREPRÉSENTATIVES

4.1 Ancienne conception

Dans les strates urbaines, les U.P.E. ont été formées de centres urbains. On a parfois combiné des centres petits et peu éloignés, mais sans considération d'optimalité des caractéristiques. Le tableau 4 donne les indices de délimitation moyens pour les U.P.E. dans les strates rurales, mixtes et urbaines. Pour les variables non-géographiques, l'indice le plus bas représente la meilleure délimitation, tandis que c'est le contraire pour les centroïdes. Les résultats sont évidemment meilleurs, au point de vue optimalité des caractéristiques, pour les strates rurales et mixtes, où on a utilisé l'algorithme de classification. Les indices élevés des centroïdes révèlent que les U.P.E. sont bien compactes.

Les unités autoreprésentatives de l'ancien plan de sondage correspondaient aux villes qui étaient assez grandes pour avoir comme rendement prévu un échantillon suffisant pour être traité par un interviewer. La limite inférieure de la taille des U.A.R. variait de 10,000 personnes dans les provinces de l'Atlantique à 29,000 personnes au Québec et en Ontario. Les grandes U.A.R. étaient stratifiées géographiquement en regroupant de 3 à 5 secteurs de recensement (S.R.) contigus, sans recherche d'optimalité. Les S.R. sont des unités géostatistiques comprenant de 3,000 à 5,000 habitants, et dont la stabilité d'un recensement à l'autre en fait des unités opérationnelles pratiques. On s'attendait à ce que ces strates soient efficaces pour l'estimation des caractéristiques, et que leur petite taille (entre 10,000 et 15,000 personnes) permette la mise à jour de l'échantillon dans les secteurs expérimentant une croissance rapide, sans d'autre part déranger l'échantillon.

En plus de la base de sondage aréolaire, une base ouverte était tenue à jour pour les immeubles d'appartements dans les grandes villes.

4.2 Étude sur la stratification

On a considéré trois grandes U.A.R. dans cette étude, soient Québec, Ottawa et Toronto. L'unité de stratification choisie a été le secteur de recensement. À cause de contraintes opérationnelles imposées par le programme de stratification, on a dû séparer Toronto en six parties correspondant généralement aux grandes divisions naturelles de la ville. Toute stratification a été menée séparément dans chacune de ces parties. On a utilisé les mêmes 16 variables de stratification qui ont finalement été choisies dans la partie N.A.R.

Deux options principales ont été évaluées:

Option 1: Stratification à deux niveaux:

- strates primaires contiguës et compactes, avec un poids de 3 sur les centroïdes et une taille déterminée en vue d'un rendement d'environ 150 logements.
- strates secondaires - 4 ou 5 par strates primaires, formulées sans contraintes géographiques.

Option 2: Stratification compacte formulée avec l'emploi de centroïdes (poids de 3) sans utiliser les vecteurs de contiguïté, avec une taille comparable à celle des strates secondaires de l'option 1.

Le tableau 5 montre les résultats de la comparaison entre l'ancienne stratification et les deux options étudiées. Comme dans la partie N.A.R., la stratification a été effectuée avec les données du recensement de 1971, pour ensuite être évaluée avec celles du recensement de 1981. On voit que les deux options étudiées donnent toujours de meilleurs indices que l'ancienne stratification, sauf peut-être pour les trois premières variables dont l'importance est de toute façon faible dans les villes. L'ancienne stratification est quand même très valable si l'on considère qu'elle a été faite sans souci d'optimalité.

1. On détermine le nombre de parties de centres urbains que recevra en moyenne une U.P.E. (N). Ce nombre dépend de la proportion de la population urbaine dans la strate ainsi que du nombre d'unités urbaines. En pratique, il a été fixé à 1 ou 2. Certaines strates ne présentent pas une population ou un nombre d'unités urbaines suffisantes se sont vues reclassifiées comme étant des strates entièrement rurales.
2. On détermine le nombre de parties par lesquelles chaque centre urbain sera divisé. Le nombre total de parties doit égaier N fois le nombre d'U.P.E. et chaque centre urbain est divisé en un nombre de parties proportionnel à sa population.
3. On applique le programme de stratification optimale en considérant chaque partie de centre urbain comme une unité de stratification distincte et en ajoutant la variable "population urbaine" aux autres variables de stratification. On ajuste le poids accordé à cette variable pour obtenir une répartition rural/urbain la plus égale possible dans chaque U.P.E., en essayant de ne pas trop affecter la compacité et l'optimisation globale. Ceci est réalisé par tâtonnement seulement. On a trouvé en pratique qu'un poids de 10 ou 15 sur la population urbaine, relativement aux autres variables, menait à des résultats satisfaisants.

Tableau 4

Indices moyens de délimitation des U.P.E.

Variables	Type de strate		
	Rurale	Mixte	Urbaine

Agriculture	8.1	8.3	9.0
Foresterie	21.8	24.5	35.9
Mines	20.6	36.0	57.0
Manufacture	15.1	22.9	53.3
Construction	9.0	11.4	22.7
Transport	9.9	12.8	22.7
Services	9.4	12.8	29.1
Employés	7.7	10.2	23.6
Chômeurs ^a	13.6	14.2	18.6
Revenu	8.9	11.2	23.7
Population 15-24	9.4	13.4	29.8
Population 55 +	7.4	13.9	34.5
Ménages 1 personne	5.1	7.4	13.0
Ménages 2 personnes	7.9	11.9	28.1
Logements possédés	6.8	12.5	29.4
Loyer brut total	5.1	7.7	14.4
Éducation secondaire	9.1	10.5	17.4
Population totale ^a	3.2	4.0	10.5
Logements ^a	5.9	8.9	18.6
Centroïde 1	91.6	92.7	99.2
Centroïde 2	90.5	91.7	97.2

^a Non utilisé comme variable dans l'optimisation.

Il y a cependant un conflit entre la compacité désirée des U.P.E. et leur hétérogénéité, à cause de la tendance des unités adjacentes à posséder des caractéristiques similaires. Étant donné les coûts d'ordinateur peu élevés, on a fait 3 délimitations par strate avec des poids sur les centroïdes de 10, 15 et 20, relativement aux autres variables. On a ensuite montré les résultats de chaque délimitation sur un graphique dont les axes représentent les centroïdes (voir figure 1). On a fait un choix entre les 3 délimitations en tenant compte de la qualité de l'optimisation des variables, reflétée par les indices de stratification, et en consultant les graphiques. On tenait également compte d'un indice de compacité. En pratique, on a choisi la plupart du temps un poids de 10 ou de 15 sur les centroïdes.

La formation des U.P.E. dans les strates mixtes a amené une contrainte supplémentaire. On désirait en effet que la proportion de la population urbaine soit à peu près la même dans chaque U.P.E. Étant donné qu'on voulait de plus avoir des U.P.E. de populations totales approximativement égales, il a donc été nécessaire parfois de partager les grands centres urbains parmi plus d'une U.P.E. La solution retenue a été la suivante:

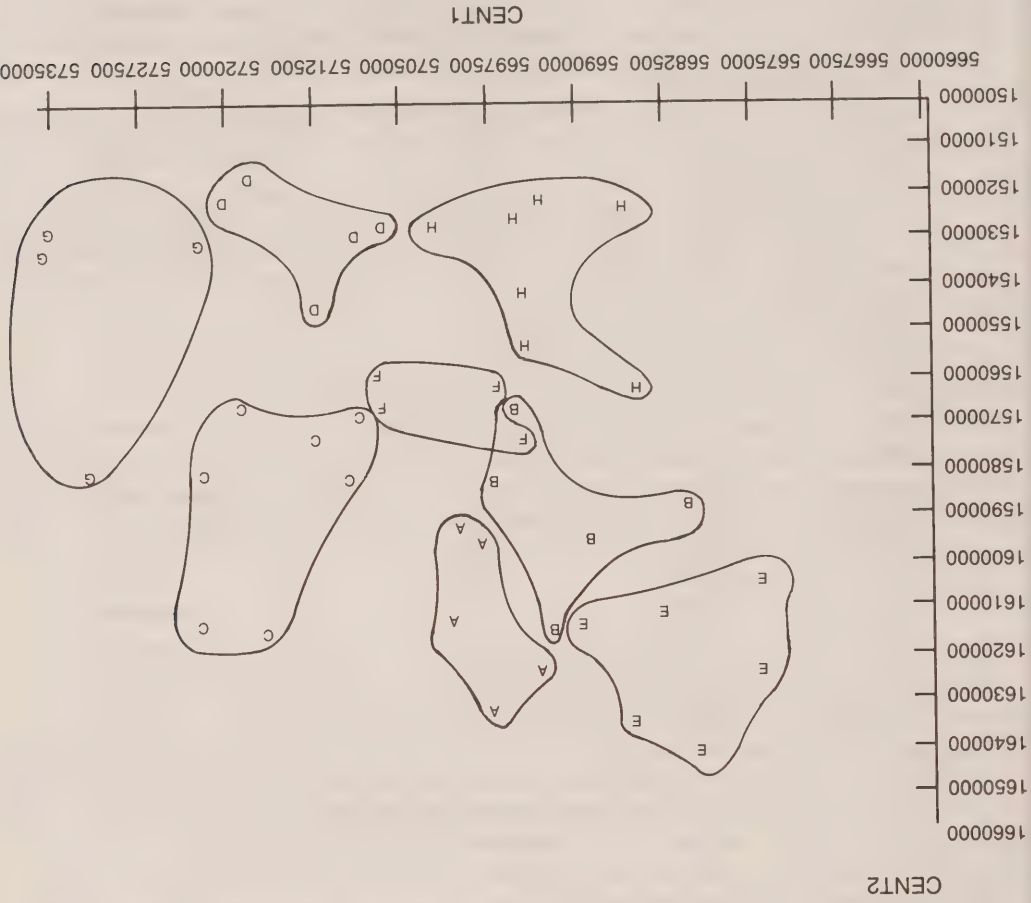


Figure 1. Exemple de délimitation d'U.P.E. Chaque unité de stratification est représentée par une lettre qui identifie l'U.P.E. à laquelle appartient l'unité. On a encercle les U.P.E. pour mieux les différencier.

Tableau 3
Indices de stratification selon les contraintes géographiques

Région économique	No. de strates	Contiguïté et centroïdes (poids de 3)	Contiguïté et centroïdes (poids de 3)	Aucune
520	2	32.2	28.5	30.2
540	3	21.8	14.1	24.9
580	4	22.8	18.9	41.4
1971	1981	1971	1981	1971
38.5	43.7	35.2	33.7	38.5

3.5 Formation des U.P.E.

On a modifié l'algorithme de classification pour effectuer la formation des U.P.E. dans les strates rurales et mixtes. Dans les strates rurales en particulier, la formation des U.P.E. ressemble beaucoup à la stratification, au point de vue conception. La seule différence tient au fait que dans la stratification, on veut minimiser la somme des carrés des variables géographiques et non-géographiques à l'intérieur de chaque strate, tandis que dans la formation des U.P.E. on veut minimiser la somme des carrés des variables géographiques (de façon à obtenir des U.P.E. compactes pour réduire les coûts) et maximiser celle des variables non-géographiques. Ce dernier critère permet d'obtenir les U.P.E. les plus hétérogènes possible en ce qui a trait aux caractéristiques, de façon à ce qu'elles soient toutes bien représentatives de la strate au moment de l'échantillonnage.

L'étude a porté sur trois régions économiques de l'Ontario, les R.E. 520, 540 et 580 (selon la numérotation de 1981). Pour chacune de ces régions, on a comparé les résultats de la nouvelle stratification (choisie pour le remaniement de l'E.P.A.) qui consiste en des strates contiguës, avec une stratification sans contrainte de contiguïté. On a stratifié en utilisant les données de 1981, et on a fait l'évaluation avec les données de 1971. Pour les strates contiguës on a utilisé les vecteurs de contiguïté avec centroïdes, tandis que pour les strates non-contiguës on a essayé deux options donnant des poids de 0 et 3 aux centroïdes, respectivement. Les variables de stratification utilisées ont été les mêmes 16 variables que l'on a déjà décrites (soit l'option 4 modifiée).

Les résultats se retrouvent dans le tableau 3. On voit qu'en général l'indice global calculé lors de la stratification est plus élevé pour les deux options où la contiguïté n'est pas nécessaire, comme on pouvait s'y attendre (colonne 1981). Cependant, ces deux options donnent également des indices plus élevés au fil du temps (colonne 1971).

A-t-on vraiment besoin de strates contiguës? Pour répondre à cette question, on devrait entreprendre une étude plus approfondie comportant des R.E. de plusieurs provinces. L'évaluation de la robustesse de la stratification poserait alors certains problèmes. Il est facile d'évaluer la robustesse en Ontario car la stratification y est faite au niveau des sous-divisions de recensement dont les frontières n'ont à peu près pas changé depuis 1971. À l'opposé, lorsque la stratification est faite au niveau des secteurs de dénombrement, qui ont des frontières changeantes d'un recensement à l'autre, il est très difficile d'obtenir des chiffres comparables entre recensements en ce qui concerne la robustesse, surtout lorsque les strates ne sont ni compactes ni contiguës.

S'il devait s'avérer que la stratification sans contiguïté est plus optimale, cela pourrait compenser pour les problèmes éventuels rencontrés dans l'estimation pour les petites régions. Cela pourrait également ouvrir de nouveaux horizons: en effet, une fois libérés des contraintes de contiguïté, pourquoi ne pas former tout d'abord les U.P.E., compactes mais pas nécessairement contiguës, pour les regrouper seulement ensuite en des strates? Encore une fois, on ne peut répondre qu'en procédant à de nouvelles études plus approfondies.

3.2.2 Etude sur la contiguïté

Comme mentionné auparavant, on a décidé de retenir l'idée des strates contiguës pour l'E.P.A. Des telles strates devraient être meilleures pour la production d'estimations pour les petites régions, parce que l'échantillon sera bien dispersé au point de vue géographique. De plus on a pensé que les strates contiguës préserveraient davantage l'efficacité du plan d'échantillonnage sur une longue période de temps.

La question a été ensuite de savoir comment utiliser les centroïdes ou les vecteurs de contiguïté, ou les deux ensemble, pour obtenir des strates compactes et contiguës sans que les contraintes géographiques ne dominent trop la minimisation des autres variables.

L'étude a été réalisée avec les mêmes 11 régions économiques. Comme prévu, l'utilisation des seuls vecteurs de contiguïté a résulté en des strates contiguës, mais montrant des configurations souvent irrégulières. D'un autre côté, l'utilisation des centroïdes seulement, même si des poids très élevés leur ont été associés, n'a donné aucune garantie de contiguïté absolue. En faisant varier les poids des centroïdes, comparativement aux autres variables, on a trouvé que l'utilisation des centroïdes avec poids égal à 3 et des vecteurs de contiguïté ont fourni un bon compromis entre la compacité et l'optimisation non-géographique.

3.3 Nouvelle conception: stratification

À la lueur des résultats décrits précédemment et aussi en tenant compte des résultats supérieurs d'un plan d'échantillonnage utilisant une stratification rural/urbain tirés d'une étude sur les variances et coûts (Choudhry, Lee et Drew 1985), on a décidé que dans la mesure du possible, on devait procéder à une stratification séparée pour toutes les régions économiques. Des restrictions s'appliquent cependant lorsque dans une R.E. la population rurale, d'une part, ou la population urbaine, d'autre part, n'est pas suffisante pour former au moins une strate. On a déterminé qu'une strate devait pouvoir donner un échantillon d'au moins 90 logements, correspondant à la sélection de deux U.P.E. d'un rendement minimal de 45 logements chacune. Dans le cas où cette contrainte ne pouvait être respectée, on a choisi de procéder à une stratification globale et donc de former des strates mixtes formées de S.D. ruraux et urbains. Ce critère a mené à l'adoption des strates séparées dans plus de 2/3 des R.E.

En ce qui concerne les variables de stratification, on a choisi comme compromis une stratification basée sur les 15 variables de l'option 4 plus *emploies*. On a ajouté *emploies* parce que son inclusion dans l'option 4, comparativement à l'option 5, a amélioré la performance des deux caractéristiques *emploies* et *revenu*. Suivant la même logique on a exclu *chômeurs* comme variable de stratification.

Pour les contraintes géographiques, on a décidé d'utiliser les vecteurs de contiguïté en combinaison avec un poids uniforme de centroïdes égal à 3 dans toutes les régions économiques. On devait également prendre une décision à propos du nombre de strates par R.E.. En pratique, dans la plupart des cas on n'a pas eu le choix. Selon le plan d'échantillonnage, chaque U.P.E. correspond à une tâche d'interviewer, et on désirait sélectionner au moins deux U.P.E. par strate afin de permettre l'estimation de la variance sans biais. Etant donné ces contraintes, dans près de 2/3 des cas on n'a formé qu'une seule strate avec 2 ou 3 U.P.E. sélectionnées, dans les parties urbaines, rurales ou une combinaison des deux. Dans les autres cas, on a stratifié de façon à sélectionner également 2 ou 3 U.P.E. par strate. On s'est basé sur une autre étude montrant de légères réductions dans la variance pour cette façon de procéder, comparativement à l'ancien plan de sondage où on sélectionnait de 4 à 6 U.P.E. par strate (Choudhry, Lee, et Drew 1985).

3.4 Etude sur la robustesse des strates contiguës et non-contiguës

Des strates robustes sont des strates qui maintiennent l'efficacité du plan d'échantillonnage au fil du temps. Après le remaniement on a entrepris une étude pour voir si les strates contiguës seraient plus robustes, comme on en a fait l'hypothèse.

Tableau 2
Indices de stratification pour les 5 options

Variables de stratification							
Global	1971	1981	1971	Global	1971	1981	Rural/urbain
Chômeurs							
7 industries	5.4	0.1	9.9	7 industries	3.8	3.4	3.8
7 industries + revenu + employés + chômeurs	5.2	2.3	10.2	7 industries + revenu + employés + chômeurs	3.4	3.4	3.4
7 industries + revenu + employés + chômeurs × 2	7.4	2.2	10.2	7 industries + revenu + employés + chômeurs × 2	5.3	5.3	5.3
17 variables	6.3	6.4	11.3	17 variables	4.7	4.7	4.7
15 variables (sauf employés + chômeurs)	3.6	0.1	9.8	15 variables (sauf employés + chômeurs)	9.0	9.0	9.0
Employés							
7 industries	2.9	0.5	8.9	7 industries	4.8	3.2	4.8
7 industries + revenu + employés + chômeurs	8.8	2.7	8.6	7 industries + revenu + employés + chômeurs	3.2	3.2	3.2
7 industries + revenu + employés + chômeurs × 2	9.1	2.8	13.1	7 industries + revenu + employés + chômeurs × 2	2.2	2.2	2.2
17 variables	14.1	7.8	12.2	17 variables	6.4	6.4	6.4
15 variables (sauf employés + chômeurs)	6.3	1.6	11.4	15 variables (sauf employés + chômeurs)	3.7	3.7	3.7
Revenu							
7 industries	7.4	5.7	18.9	7 industries	9.5	5.9	9.5
7 industries + revenu + employés + chômeurs	11.2	6.8	22.1	7 industries + revenu + employés + chômeurs	9.5	5.9	9.5
7 industries + revenu + employés + chômeurs × 2	10.3	6.8	28.3	7 industries + revenu + employés + chômeurs × 2	9.5	5.9	9.5
17 variables	10.5	9.4	24.4	17 variables	11.9	11.9	11.9
15 variables (sauf employés + chômeurs)	21.0	5.3	28.9	15 variables (sauf employés + chômeurs)	4.5	4.5	4.5
Agriculture							
7 industries	7.4	9.7	37.0	7 industries	26.0	28.7	26.0
7 industries + revenu + employés + chômeurs	7.6	7.8	40.0	7 industries + revenu + employés + chômeurs	28.7	28.7	28.7
7 industries + revenu + employés + chômeurs × 2	8.6	7.9	43.2	7 industries + revenu + employés + chômeurs × 2	31.0	31.0	31.0
17 variables	6.1	1.1	40.3	17 variables	31.8	31.8	31.8
15 variables (sauf employés + chômeurs)	7.0	0.4	42.7	15 variables (sauf employés + chômeurs)	29.0	29.0	29.0
Manufacture							
7 industries	14.7	8.5	16.9	7 industries	13.2	12.1	13.2
7 industries + revenu + employés + chômeurs	10.9	6.6	16.5	7 industries + revenu + employés + chômeurs	12.1	12.1	12.1
7 industries + revenu + employés + chômeurs × 2	5.5	4.3	14.8	7 industries + revenu + employés + chômeurs × 2	16.1	16.1	16.1
17 variables	12.5	13.5	13.3	17 variables	10.7	10.7	10.7
15 variables (sauf employés + chômeurs)	7.2	1.4	14.1	15 variables (sauf employés + chômeurs)	16.4	16.4	16.4

Tableau 1

Options de stratification selon les variables

Variables	Option de stratification				
	1	2	3	4	5

Industries (7) ^a	x	x	x	x	x
Revenu		x	x	x	x
Employés		x	x		x
Chômeurs		x	x ^b		x
Démographie (2) ^c			x	x	x
Logement (4) ^d			x	x	x
Éducation (1) ^e			x	x	x

^a nombre de personnes employées dans les secteurs agriculture, forêt et pêche, mines, manu-
facture, construction, transport, services.
^b poids double pour nombre de chômeurs.
^c population 15-24 ans, population 55 ans et plus.
^d ménages à 1 personne, ménages à 2 personnes, logements possédés, loyer brut total.
^e personnes ayant une éducation secondaire.

On a utilisé une filière de conversion entre les unités géographiques des deux recensements, pour faire l'évaluation basée sur le recensement de 1981. Les indices basés sur les données de 1981 ont été jugés plus au point pour l'évaluation, étant donné qu'en réalité l'âge moyen des données de stratification sera de 7 ou 8 ans pendant la vie du plan de sondage. On retrouve dans le tableau 2 les indices basés sur les deux recensements.

Pour cette étude on a choisi de former des strates contiguës et compactes, en utilisant les vecteurs de contiguïté et les variables centroïdes avec des poids moyens de trois (voir sous-section 3.2.2). Le nombre de strates à former par R.E. était le même pour toutes les options. On a tiré les conclusions suivantes des résultats du tableau 2:

Type de Stratification: La stratification rural/urbain était de beaucoup supérieure à la stratification globale dans le cas de la variable *agriculture*, ce qui n'est pas étonnant. Le même phénomène s'est produit pour la variable *manufacture*, quoique moins spectaculaire. Pour la variable *revenu* la stratification rural/urbain était également meilleure, au départ mais elle n'était pas très robuste (i.e., l'indice se détériore avec le temps). La stratification rural/urbain était meilleure pour la variable *chômeurs*, tandis qu'il n'y avait pas beaucoup de différence pour *employés*.

Variables de stratification: L'option 4, en combinaison avec la stratification rural/urbain était nettement supérieure pour la variable *chômeurs*. En ce qui concerne les autres variables, l'option 5 était légèrement meilleure que les autres pour *employés* et *revenu*.

À l'intérieur de chaque strate, de 12 à 15 U.P.E. ont été formées, de façon à être semblables à la strate relativement aux variables de stratification, et relativement au rapport entre la population rurale et urbaine. Les parties rurales des U.P.E. ont été formées de S.D. continents, et les parties urbaines ont été choisies de façon à être aussi près géographiquement que possible de la partie rurale. Les tailles des strates et des U.P.E. ont été déterminées de sorte qu'avec deux U.P.E. sélectionnées par strate, la taille de l'échantillon prévue soit équivalente à une tâche d'interviewer. Suivant ces critères, selon la province, la population des U.P.E. variait entre 3,000 et 5,000 personnes. À l'intérieur des U.P.E., l'échantillonnage a été fait en 2 ou 3 étapes.

3.2 Les études sur la stratification au moment du remaniement

Le but de nos études était de tirer des conclusions permettant de prendre des décisions relatives aux aspects suivants de la stratification: variables à utiliser, type de strates (entières-ment rurales, entièrement urbaines ou mixtes), et importance à accorder à la contiguïté. Étant donné le temps très limité pour les études avant le moment de la formation des nouvelles strates et U.P.E., et l'attente générale que des strates formées d'unités contiguës seraient préférables au fil du temps à des strates formées d'unités non-contiguës, les deux premiers aspects ont été jugés prioritaires.

En ce qui concerne la contiguïté, on a dû expérimenter pour trouver le meilleur moyen de l'atteindre, soit par les vecteurs de contiguïté, les centroïdes ou une combinaison des deux. Cependant, après le remaniement, on a entrepris une étude plus approfondie examinant le choix entre des strates contiguës et des strates non-contiguës.

3.2.1 Étude sur les variables et le type de stratification

Une contrainte de la méthode de stratification utilisée dans l'ancien plan de sondage était le nombre limité de variables de stratification que l'on pouvait considérer (3 par R.E.). Avec le nouvel algorithme, on s'est libéré de cette contrainte. En plus des sept variables d'industrie, on a voulu voir l'effet causé par l'utilisation de variables reliées au sujet de l'enquête, telles que: taux de personnes occupées, taux de chômage et niveau de revenu, ainsi que par des caractéristiques telles que: niveau d'éducation, type de logement et nombre total d'individus. Ces dernières caractéristiques se sont révélées très efficaces dans des études semblables menées par le U.S. Bureau of the Census pour le Current Population Survey.

Le tableau 1 décrit les diverses options étudiées en ce qui a trait au choix des variables. En ce qui concerne le type de stratification, on a décidé d'étudier l'effet de former des strates séparément pour les parties rurales et urbaines des R.E., comme une alternative à la méthode mixte de l'ancien plan.

Les contraintes du plan de sondage selon lesquelles les U.P.E. doivent avoir des populations presque égales, tandis que le rapport entre la population rurale et la population urbaine doit demeurer presque le même pour chaque U.P.E., ont résulté par le passé en un manque de contiguïté entre les parties rurales et urbaines des U.P.E. Ceci a mené à une érosion de la présomée correspondance entre l'U.P.E. et la tâche d'interviewer. On a pensé que les strates séparées en parties rurales et urbaines, qui pourraient être sous-stratifiées d'une façon optimale, constituaient une solution possible à ce problème.

L'étude a porté sur 11 régions économiques réparties à travers tout le Canada. On a créé des strates en utilisant des données du recensement de 1971, et on les a évaluées en utilisant les données du recensement de 1981. Pour faire la stratification, on a choisi comme unité de stratification les secteurs de dénombrement du recensement de 1971, sauf au Québec et en Ontario. Pour ces deux provinces on a choisi les sous-divisions de recensement, car le grand nombre de S.D. dans certaines R.E. (jusqu'à 400) aurait occasionné des coûts d'exécution des programmes informatiques trop élevés.

La deuxième condition est plus difficile à vérifier. Le principe est qu'une strate est dite contiguë si chaque paire d'unités de cette strate peut être reliée par une chaîne continue d'unités provenant de cette même strate. Soit l'unité j que l'on veut déplacer de la strate A à la strate B . Il est donc nécessaire de chercher, pour chaque paire d'unités du vecteur de contiguïté de l'unité j appartenant à la strate A , un autre lien parmi les unités de la strate A . À ce stade, le problème se compare à celui de trouver son chemin dans un labyrinthe.

Un algorithme a aussi été conçu pour créer de façon aléatoire des partitions initiales contiguës.

2.4 Pondération des variables

Les facteurs de pondération revêtent une importance toute particulière. Ce sont eux qui établissent l'apport de chaque variable à la classification.

Il est habituellement préférable de standardiser les variables, en rendant les facteurs de pondération inversement proportionnels à la somme des carrés totale de chaque variable. Cette standardisation permet d'obtenir un apport comparable de chaque variable à la classification.

Si après la standardisation on veut qu'une ou plusieurs variables se voient accorder plus d'importance relativement aux autres variables dans l'optimisation, on peut le faire en spécifiant un poids supérieur à 1 (la normale). Par exemple, une variable avec un poids de 2 aurait une importance double. Comme décrit dans la section 3.2 on a dû essayer plusieurs combinaisons de poids pour les variables à caractère géographique et non-géographique dans le but d'obtenir des strates compactes sans trop affecter la minimisation des autres variables.

3. LA STRATIFICATION DANS LES UNITES NON AUTOREPRESENTATIVES

3.1 Ancienne conception (Platek et Singh 1976)

Pour l'E.P.A., chacune des dix provinces canadiennes est divisée en un certain nombre de régions économiques (R.E.), composées de secteurs ayant des structures économiques semblables. Les frontières des R.E. sont déterminées en consultation avec les provinces. Ces R.E. sont considérées comme des strates primaires. L'étape de stratification suivante est la partition de chaque R.E. en unités autoreprésentatives (U.A.R.) et unités non autoreprésentatives (U.N.A.R.). Les unités autoreprésentatives sont les villes où l'échantillon prévu est assez important pour constituer au moins une tâche d'interviewer; la partie N.A.R. comprend le reste de la R.E. Des plans d'échantillonnage différents sont suivis dans les U.A.R. et U.N.A.R., parce que la population dans les U.N.A.R. est beaucoup plus dispersée, rendant nécessaire un plus grand nombre d'étapes d'échantillonnage. Pour ces mêmes raisons, on retient le concept des U.A.R. et U.N.A.R. dans le remaniement.

Dans l'ancien plan, on a stratifié la partie N.A.R. de chaque R.E. en un maximum de 5 strates contiguës ayant une population entre 36,000 et 75,000 personnes, par rapport aux caractéristiques principales de la population du recensement de 1971, comme décrit ci-dessous, et comme élaboré en plus de détail par Platek et Singh (1976).

La population active a été répartie en 7 catégories selon l'industrie. Dans chaque R.E., on a choisi les 3 catégories les plus importantes selon certains critères spécifiques. Les municipalités groupées, qui représentent les régions géographiques comprises dans une municipalité rurale, et qui, par conséquent, renferment souvent des municipalités urbaines géographiquement plus petites, ont été les unités de stratification. En comparant, pour chacune de ces unités, les proportions de la population active appartenant à chacune des trois catégories, avec les proportions correspondantes au niveau de la R.E., on a pu identifier les unités montrant une certaine ressemblance entre elles pour les regrouper dans des strates. Cette comparaison a été effectuée visuellement à l'aide de graphiques. En général, il a fallu procéder à certains rajustements pour satisfaire les exigences relatives à la taille et à la contiguïté des strates.

où la valeur de la fonction objective est encore plus faible. Pour aller au-delà du minimum local, Friedman et Rubin décrivent deux procédures, soient le "forcing pass" et le "reassignment pass". En appliquant leur algorithme aux données décrites dans leur article, ils obtiennent la plus haute valeur connue de la fonction objective 10 fois sur 14 tentatives en utilisant différentes partitions initiales (ils utilisent une autre fonction objective qui est maximisée). Avec des données moins bien structurées, la plus haute valeur est atteinte 3 fois sur 11, sans toutefois l'assurance d'avoir obtenu la solution optimale. À leur avis, les méthodes du "forcing pass" et du "reassignment pass" ne sont utiles qu'à l'occasion. Ils ont plus confiance aux résultats obtenus en utilisant plusieurs partitions initiales. Ce point de vue est soutenu par Judkins et Singh (1981). On a donc décidé d'utiliser la technique de plusieurs partitions initiales.

Parce que l'algorithme ne déplace qu'une unité à la fois, le calcul de la fonction objective en est simplifié. Après le calcul initial de la fonction objective, il suffit de recalculer la contribution à la fonction objective des deux groupes impliqués dans le transfert de l'unité en jeu.

2.3 La contiguïté

Dans les plans de sondage antérieurs de l'E.P.A., on avait adopté des strates formées d'unités géographiques contiguës, c'est-à-dire que chaque unité d'une strate donnée devait toucher à au moins une autre unité de la même strate. Une des raisons principales était une présomption que de telles strates préserveraient l'efficacité du plan d'échantillonnage sur une plus longue période de temps que dans le cas des strates formées d'unités non-contiguës. En vue d'évaluer cette présomption et d'adopter la meilleure stratification possible, nous avons considéré deux moyens de tenir compte de la géographie dans la stratification. La première méthode est élaborée par Dahmström et Hagnell (1978), et consiste en l'utilisation de centroïdes comme variables d'intérêt. Cette méthode fait appel à deux variables géographiques (centroïdes) qui sont des transformations de longitude et de latitude. Cette méthode donne des strates compactes, c'est-à-dire des strates où la distance entre les unités est rendue minimale par la minimisation de la somme des carrés intra-groupe usuelle des centroïdes. Cette minimisation est tempérée par la minimisation des autres variables d'intérêt. Aussi il n'y a pas de garantie que les strates soient ainsi formées d'unités contiguës.

L'autre méthode, que nous appelons l'approche de vecteurs de contiguïté, est nouvelle. Elle garantit des strates contiguës, mais pas nécessairement compactes. Des études décrites dans la section 3, ont porté sur l'utilisation de l'une ou l'autre des méthodes seulement ou d'une combinaison des deux méthodes.

2.3.1 Vecteurs de contiguïté

Afin d'assurer la formation de strates contiguës, on a procédé de la façon suivante. On fait l'optimisation comme décrite dans la section précédente mais en commençant dans ce cas avec une partition initiale qui est contiguë, et en permettant le déplacement de l'unité "j" de la strate A à la strate B, seulement si en plus de réduire les sommes des carrés, les conditions suivantes sont respectées:

- (i) l'unité j est contiguë à une unité de la strate B
- (ii) le déplacement de l'unité j à la strate B ne dérangera pas la contiguïté de la strate A.

Pour vérifier ces deux conditions, il est essentiel de connaître les liens de contiguïté entre les unités. Par conséquent, à chaque unité doit être assigné un vecteur de contiguïté qui contient la liste des unités qui y sont contiguës.

La première condition est facile à vérifier. Pour s'assurer que l'unité "j" est contiguë à une unité de la strate B, il suffit de trouver dans son vecteur de contiguïté une unité qui appartient à la strate B.

$$SCB_i = \sum_{k=1}^L \frac{T_k}{T_{..}} \left(\frac{T_k}{T_{..}} X_k - {}^iX_{k..} \right)^2.$$

Leurs sommes des carrés pondérées sur toutes les variables sont données respectivement par

$$SCW = \sum_{i=1}^p w_i SCW_i$$

et

$$SCB = \sum_{i=1}^p w_i SCB_i$$

La somme des carrés intra-groupes de la variable i , SCW_i , est aussi l'expression de la variance de l'estimateur de ${}^iX_{..}$ lorsque une strate est choisie avec PPT et subséquemment une unité de cette strate est choisie avec PPT.

Comme d'habitude, nous avons le résultat suivant:

$$SCT_i = SCW_i + SCB_i, (i = 1, \dots, p)$$

et

$$SCT = SCW + SCB.$$

La fonction objective du programme de stratification se trouve à être SCW , la somme des carrés intra-groupes pondérée sur toutes les variables. Elle doit être minimisée. On définit l'indice de stratification associé à la variable i , I_i , comme étant:

$$I_i = 100 \times \frac{SCB_i}{SCT_i} \quad i = 1, \dots, p$$

Une valeur élevée de l'indice indique une bonne stratification.

2.2 Recherche de la meilleure classification

Afin d'identifier la meilleure classification, une méthode consisterait à générer toutes les partitions possibles de N unités en L groupes. Il suffirait ensuite de retenir celle qui minimise la fonction objective. Cette tactique est rarement faisable puisque le nombre de partitions possible peut être exagérément grand.

Friedman et Rubin (1967) suggèrent l'algorithme suivant. Commençons d'abord avec une partition quelconque des N unités en L groupes. Considérons maintenant le transfert d'une unité à un groupe autre que celui auquel elle appartient. Cette unité sera transférée au groupe qui apportera la plus forte réduction de la fonction objective. Toutefois, l'unité demeurera dans son groupe original si aucun transfert n'apporte de réduction. En se servant maintenant de la partition ainsi engendrée, traitons la deuxième unité de la même manière, ensuite la troisième, jusqu'à la $N^{\text{ième}}$. L'application de cette procédure à chaque unité devient une itération que les auteurs appellent "hill-climbing pass". Après plusieurs itérations, l'algorithme atteint un point où aucun transfert de quelque unité que ce soit ne permet de réduire la fonction objective. Ce point est dit un minimum local de la fonction objective parce qu'il dépend de la partition initiale utilisée. Une autre partition initiale aurait pu atteindre un autre point

2. ALGORITHME DE STRATIFICATION

L'algorithme de base utilisé pour la stratification est un algorithme multi-varié, non-hiérarchique élaboré par Friedman et Rubin (1967). Ce choix repose sur les résultats d'études faites par Judkins et Singh (1981) et Kostanich, Judkins, Singh et Schantz (1981), qui ont évalué plusieurs algorithmes de stratification pour le Current Population Survey du U.S. Bureau of the Census.

Ces derniers ont modifié la fonction objective de l'algorithme dans le cas de l'échantillonnage avec probabilité proportionnelle à la taille (PPT), et nous avons ajouté la capacité de formuler des strates compactes et contiguës. Une description plus complète de ce qui suit se retrouve dans Foy (1984).

2.1 La fonction objective de l'algorithme

L'algorithme vise à partitionner les unités de stratification (secteurs de dénombrement du recensement) dans des strates les plus homogènes possible à l'égard de plusieurs variables d'intérêt, c'est-à-dire en minimisant les sommes des carrés à l'intérieur de chaque strate. Les expressions pour les sommes des carrés dans le cas de l'échantillonnage avec PPT suivent après l'introduction de la notation utilisée:

L = nombre de strates à former,

N = nombre total d'unités (secteurs de dénombrement),

N_k = nombre d'unités dans le groupe (strate) k ; ($N_1 + N_2 + \dots + N_L = N$),

T_{jk} = mesure de la taille de l'unité j du groupe k ,

$T_{..k}$ = mesure de la taille du groupe k ,

$T_{..}$ = taille totale,

$X_{jk}^{..}$ = valeur observée de la variable i pour l'unité j du groupe k ,

$X_{..k}^{..}$ = total des valeurs observées de la variable i dans le groupe k ,

$X_{..}^{..}$ = total des valeurs observées de la variable i ,

W_i = facteur de pondération de la variable i (voir section 2.4 pour plus de détails),

p = nombre de variables d'intérêt.

Ainsi, l'expression de la somme des carrés totale, avec PPT, de la variable i est donnée par

$$SCT_i = \sum_{k=1}^L \sum_{j=1}^{N_k} \frac{T_{jk}}{T_{..}} \left(\frac{T_{jk}}{T_{..}} X_{jk}^{..} - \frac{T_{..}}{T_{..}} X_{..}^{..} \right)^2.$$

Ceci est aussi l'expression de la variance de l'estimateur de $X_{..}^{..}$ lorsqu'une unité est choisie avec PPT. La somme des carrés totale pondérée sur toutes les variables est donc

$$SCT = \sum_{i=1}^p W_i SCT_i.$$

Quant aux sommes des carrés intra-groupes et inter-groupes, elles sont obtenues respectivement par les expressions suivantes:

$$SCW_i = \sum_{k=1}^L \sum_{j=1}^{N_k} \frac{T_{jk}}{T_{..}} \left(\frac{T_{jk}}{T_{..}} X_{jk}^{..} - \frac{T_{..}}{T_{..}} X_{..}^{..} \right)^2$$

La stratification dans l'enquête sur la population active du Canada

J.D. DREW, Y. BÉLANGER, et P. FOY¹

RÉSUMÉ

On décrit l'utilisation d'un algorithme de classification multi-variee pour effectuer la stratification pour l'enquête sur la population active. L'algorithme elabore par Friedman et Rubin (1967), est modifie de maniere a traiter la formation de strates geographiquement contigues, et a faire la delimitation d'unites primaires d'echantillonnage (U.P.E.) heterogenes mais compactes a l'interieur des strates. Des etudes portant sur les variables de stratification, la robustesse de la stratification au fil du temps, et le type de stratification sont decrites.

MOTS CLÉS: Algorithme de classification multi-variee; stratification geographique; enquete permanente.

1. INTRODUCTION

L'enquete sur la population active du Canada (E.P.A.) est remaniee apres chaque recensement decennal de la population et des logements. Dans le cadre du remaniement suivant le recensement de 1981, on a mene un programme intensif de recherche sur divers aspects du plan de sondage (Singh, Drew, et Choudhry 1984). Ce rapport decrit la partie du programme de recherche portant sur les methodes de stratification.

Etant donne que l'E.P.A. est utilisee, non seulement pour fournir de l'information sur les caracteristiques de la population active, mais aussi comme plan de sondage general pour diverses autres enquetes-menages, l'un des principaux objectifs du remaniement etait d'accroitre la flexibilite de l'E.P.A. pour des applications generales. La stratification a ete consideree comme un moyen d'ameliorer l'efficacite du plan d'echantillonnage en ce qui a trait aux applications generales, de meme que pour des variables d'interet particulier pour l'E.P.A., en adoptant des methodes plus rigoureuses que dans l'ancien plan.

On a donc decide de considerer l'utilisation d'algorithmes de classification multi-variee pour stratifier et mettre en gappes et de les comparer avec les methodes utilisees dans l'ancien plan de sondage. On a choisi un algorithme non-hierarchique elabore par Friedman et Rubin (1967), en nous fiant aux resultats d'evaluations des differents algorithmes faites par Judkins et Singh (1981) dans le cadre de remaniement du Current Population Survey du U.S. Bureau of Census. La description de l'algorithme de base, et des extensions que nous avons developpees se retrouve dans la section 2.

Les sections 3 et 4 decrivent les etudes d'evaluation et la stratification finalement adoptee dans les deux grands types de secteurs du plan de sondage de l'E.P.A., soient les Unites Non Auto Repräsentatives (U.N.A.R.) et les Unites Auto Repräsentatives (U.A.R.). La section 4 decrit egalement comment on a adapte l'algorithme pour delimitier les unites primaires d'echantillonnage a l'interieur des strates de la partie N.A.R.

On conclut dans la section 5 avec quelques observations sur la possibilite d'adapter le systeme developpe a d'autres applications.

¹ J.D. Drew et Y. Bélanger, Division des méthodes de recensement et d'enquêtes-ménages, et P. Foy, Division des méthodes d'enquête pour les entreprises, Statistique Canada, Ottawa, (Ontario), KIA 0T6.

Egalement dans cet esprit, Statistique Canada a constitué un réseau de comités consultatifs, dont un sur la méthodologie statistique.

Statistique Canada a retenu les services de plusieurs statisticiens probabilitistes comme consultants.

Je crois qu'il y a beaucoup de possibilités que pourraient exploiter Statistique Canada et les universités en organisant des ateliers mixtes et en collaborant à divers projets.

Il faudrait une meilleure compréhension mutuelle chez les praticiens et les théoriciens. Peut-être conviendrait-il de changer les critères et les normes ayant trait à la publication des données?

Peut-être faut-il modifier les bases sur lesquelles sont évaluées les demandes de subventions adressées au conseil de recherches en sciences naturelles et en génie du Canada?

Peut-être qu'il serait utile de modifier les programmes de formation. Peut-être que Statistique Canada devrait offrir un prix pour la mise au point de solutions productives dans des secteurs d'exploitation des bureaux de la statistique où le besoin s'en fait particulièrement sentir. Peut-être faudrait-il avoir une liste permanente des dix solutions les plus nécessaires; cette liste serait un stimulant pour les statisticiens probabilitistes et un moyen de rester en contact avec eux.

Peut-être la société Statistique du Canada devrait-elle faire en sorte de créer une tradition selon laquelle l'allocution qui suit le dîner de la conférence annuelle aurait toujours pour thème le rapport entre statisticiens et statisticiens?

D'une part, donc, la statistique probabiliste a produit la notion utile et importante d'estimation objective de l'erreur; le public a été initié à cette notion et s'attend à la voir appliquée. D'autre part, il y a beaucoup d'informations statistiques très utilisées qui sont produites par des gens qu'on appelle des statisticiens, mais pour lesquelles des mesures de l'erreur ne sont ni ne peuvent actuellement être données.

Théoriquement, il semble y avoir plusieurs possibilités.

a) on pourrait ne plus englober dans la science statistique, et l'activité des bureaux de la statistique et celle des statisticiens probabilistes, et abandonner le principe d'une relation entre les deux.

b) la statistique probabiliste pourrait orienter ses travaux dans le sens de la recherche d'une technique adaptée à la réalité de l'élaboration de données complexes.

c) les statisticiens pourraient s'employer à rééduquer le public de façon qu'il cesse de croire à la possibilité de mesures nettes et objectives de l'incertitude.

Dans la pratique cependant, seulement la possibilité b) peut être envisagée. De fait, elle serait plus fructueuse pour tous ceux dont la statistique est le métier.

Dans un article paru dans la revue *Science* l'an dernier, un spécialiste de la philosophie des sciences, Ian Hacking, affirmait que les statisticiens avaient discrètement transformé le monde dans lequel nous vivons, non par la découverte de faits nouveaux ou par des progrès techniques, mais en changeant notre manière de raisonner, de faire des expériences et de nous former ensuite une opinion.

Cela fait plaisir de lire un jugement celui-là sur l'importance de la statistique probabiliste telle qu'en ont jeté les bases des gens comme Fisher, Neyman, Pearson, Wald et d'autres.

Mais j'aimerais souligner, pour en revenir au thème principal de mon exposé, qu'il y a une autre catégorie de statisticiens — les praticiens des bureaux de la statistique — qui ont joué un rôle dans cette transformation discrète du monde, encore que d'une manière qui est l'opposé du point de vue de M. Hacking.

Les praticiens de la statistique découvrent de nouveaux faits, définissent de nouvelles notions et élaborent des définitions opératoires dont les applications peuvent servir au public. Ils font effectivement des découvertes techniques dans le domaine du calcul, de la diffusion électronique de l'information, de l'infographie, des systèmes de classification, des systèmes de comptabilité nationale, etc.

Encore une fois, mon intention n'est pas ici de faire, implicitement ou explicitement, un jugement sur la valeur relative de deux types d'activité. Ce que je me demande, c'est ce qu'il y a de *réel* et d'*imaginaire* dans le rapport entre la statistique pratique et la statistique théorique. La plupart des choses que font les praticiens ne découlent pas, en réalité, des constructions mathématiques, des théories et des présupposés de la statistique probabiliste, ni n'ont avec eux un lien immédiat. Et pourtant les praticiens se sentent d'une certaine façon obligés de faire semblant d'accorder de l'importance à un lien fondamental qu'auraient leurs travaux avec les concepts théoriques.

Par ailleurs, les théoriciens continuent de penser plus ou moins confusément que si de plus nombreux praticiens avaient des préoccupations théoriques, alors la statistique probabiliste pourrait avoir un effet *réel* sur les travaux des bureaux de la statistique.

Le potentiel que peut receler une collaboration plus efficace des praticiens et des théoriciens ne sera pas libéré sans un effort des uns et des autres.

Et ce n'est pas moi qui peux ce soit proposer des solutions qui auraient le caractère de révélations.

Cependant, il faut évidemment établir de meilleurs canaux de communication. Statistique Canada a créé en ce sens un programme de bourses de perfectionnement et de stage.

Mais revenons à l'évaluation objective de l'erreur, notion censément la plus importante de la statistique probabiliste. Statistique Canada ne produit pas de mesure de l'erreur d'estimation de l'indice des prix à la consommation. Nous ne publions *pas* d'estimations par intervalle de cet indice. Nous ne testons pas l'hypothèse de la stabilité de cet indice d'un mois à l'autre. Et nous ne produisons pas d'estimations composites dont la fonction serait de réduire la variance de l'erreur aléatoire.

On nous adresse de temps à autre des questions et des critiques sur cet état de choses, même par des gens qui ne sont ni des statisticiens, ni des scientifiques. Il semble que l'habitude de se voir communiquer des résultats de sondage d'opinion commence à donner au public le sentiment qu'une estimation d'erreur doit accompagner toute publication d'estimations. L'expression "19 fois sur 20" fait maintenant partie du vocabulaire de tous les lecteurs de journaux. Bien entendu, les sondages d'opinion existent depuis longtemps. George Gallup fait remonter l'usage de ce genre de sondage au moins jusqu'à 1824, année où un journal de Pennsylvanie publiait les résultats de ce qu'on avait alors appelé un "vote blanc pris sans considération de l'allégeance partisane". Les moyens de communication modernes et l'ordinateur ont fait proliférer les sondages. L'omniprésence des sondages a rendu le public plus conscient du fait qu'un statisticien (ou quiconque) peut effectuer une enquête par échantillonnage, faire des inférences et attribuer une mesure d'incertitude aux estimations.

Dans son rapport de 1983, le vérificateur Général du Canada posait la question de la mesure de la qualité des données produites par Statistique Canada. Le vérificateur recommandait que Statistique Canada définisse et diffuse un plus grand nombre de mesures de la qualité de ses données. L'agence fédérale a répondu officiellement que cette recommandation ne pouvait être appliquée intégralement vu l'impossibilité de produire des mesures de la qualité pour plusieurs genres de données, en particulier pour les données synthétiques. La réponse de Statistique Canada disait encore qu'il serait plus réaliste de donner une *description* complète des facteurs connus de baisse possible de la qualité et, quand il était possible d'en établir, des mesures de la qualité.

Certes, Statistique Canada publierait plus d'estimations d'erreurs si la chose était jugée possible. Il ne faut pas voir là une volonté de rejeter la possibilité de l'erreur. Comme disait à ses étudiants le professeur R. C. Bose, l'erreur est humaine, et les statisticiens sont humains. Toutefois, les estimations d'erreurs reposent habituellement sur des hypothèses qui simplifient à outrance la situation. Par exemple, l'enquête sur la population active utilise des échantillons de ménages, non des échantillons de particuliers, qui ont les mêmes chances d'être choisis. En outre, étant donné la nature du plan de sondage, les ménages eux-mêmes n'ont pas des chances égales de faire partie de l'échantillon. La fraction de sondage est d'environ 1 sur 125, à l'échelle nationale, et peut atteindre 1 sur 24 dans les provinces à faible population. Pouvons-nous, alors, supposer que tous les particuliers sont indépendants et ont des chances égales d'être en chômage? Les données sont recueillies au moyen d'interviews, et l'interviewer comme le répondant peut commettre une erreur par inadvertance ou délibérément. Pouvons-nous ignorer toutes les sources possibles d'erreur hors l'erreur d'échantillonnage? Les personnes composant un ménage donné font l'objet d'un sondage pendant six mois consécutifs, un ménage sur six étant chaque mois retiré de l'échantillon tandis qu'un autre y est introduit pour le remplacer. Ainsi, dans n'importe quel mois donné, les répondants n'ont pas tous rempli le questionnaire le même nombre de fois. Pouvons-nous supposer que les six réponses sont indépendantes dans le temps? Pendant les six mois de sondage, un logement peut changer d'occupants. Et, naturellement, il y a les problèmes habituels de non-réponse, d'observations aberrantes, d'erreurs d'entrée des données, de calcul, d'impression, etc. Le problème des écarts par rapport à l'hypothèse "normale" de la théorie statistique reste un sujet de constante préoccupation pour certains théoriciens et praticiens de Statistique Canada.

Pour juger si la statistique probabiliste est une bonne chose, nous devrions nous demander s'il en découle une technique permettant de produire des données utiles et valables. Au lieu de cela, les statisticiens ont tendance à faire le contraire, c'est-à-dire à se demander si les travaux des praticiens sont valables à la lumière des préceptes de la statistique probabiliste. La statistique probabiliste a produit des concepts, des modèles et des méthodes en grand nombre. Voici quelques exemples.

Prise de décision dans un contexte d'incertitude
 Probabilité subjective
 Science de l'inférence
 Inférence de la vraisemblance
 Estimation par la méthode de Bayes
 Analyse de séries chronologiques
 Tests d'hypothèses
 Tests de signification
 Estimation du niveau de confiance
 Estimation d'erreurs de sondage
 Méthodes de classification
 Analyse de régression
 Composantes de la variance
 Plan d'expérience
 Plan d'enquête par sondage
 Estimateurs sans biais

Beaucoup d'auteurs ont dit que le concept le plus fondamental de la statistique probabiliste appliquée était l'*évaluation objective de l'incertitude*.

Mais je dois vous avouer que cette notion, si séduisante qu'elle puisse être et quelle que soit sa profondeur théorique, ne correspond pas à la réalité des travaux et du mandat des bureaux de la statistique.

Qu'il me soit permis de montrer par l'exemple l'importance du travail des praticiens. Vous pouvez faire vous-mêmes une expérience. Dressez la liste des questions qui vous semblent avoir un intérêt pour la société canadienne. Votre liste touchera à des points comme l'emploi et le chômage, le revenu des personnes âgées, la condition féminine, la croissance économique, les échanges commerciaux et la balance des paiements, la formation des familles, la distribution de la population, le déficit de l'état, et ainsi de suite.

Si vous y réfléchissez, vous constaterez que, pour la majorité de ces questions, vos impressions et vos connaissances dépendent assez directement des données statistiques produites par les praticiens, *principalement* par ceux de Statistique Canada. On pourrait faire la même constatation pour n'importe quel pays.

Pour préciser encore ma pensée par un exemple précis, j'aimerais donner ici un bref aperçu des applications possibles de l'indice des prix à la consommation.

L'indice des prix à la consommation est mis à jour chaque mois par Statistique Canada d'après des relevés mensuels des prix relatifs à un panier déterminé de biens et de services. Cet indice est le plus utilisé des indicateurs du taux d'inflation. On l'appelle souvent l'indice du coût de la vie. L'indice des prix à la consommation a un effet direct ou indirect sur presque tous les canadiens. Cet indice ou l'une ou l'autre des composantes dont il est la moyenne pondérée est utilisée pour faire des calculs ou établir des définitions dans les domaines suivants: impôt, conventions collectives, allocations familiales, pensions de sécurité de la vieillesse, contrats de location, assurances, pensions alimentaires, paiements pour enfants à charge, paiements pour enfants d'anciens combattants, remboursement de prêts étudiants. Il entre aussi en ligne de compte dans beaucoup d'autres formes de contrats et de textes de réglementation.

Mais ce que *je cherche à faire*, c'est analyser la relation qu'il peut y avoir entre ces deux formes d'activité qu'on appelle *statistique* dans un cas comme dans l'autre et qui sont l'occupation de spécialistes qu'on appelle tous des *statisticiens*.

Evidemment, nous pourrions nous contenter de dire qu'il s'agit d'un simple cas de polysémie, qui est le fait pour un mot d'avoir plus d'un sens. Ou nous pourrions tout simplement ignorer le problème. Mais cela ne serait ni sage, ni très positif.

Vous connaissez tous l'ouvrage classique de Kendall et Stewart sur la théorie de la statistique. Le premier volume a 396 pages de texte sans compter les tableaux et l'index. Ces 396 pages parlent de concepts théoriques de la statistique probabiliste et de la construction mathématique de formules diverses.

L'ouvrage commence par le passage suivant de O. Henry que citent les auteurs.

«Asseyons-nous sur ce tronc d'arbre, dis-je, et oublions l'inhumanité et les ribauderies des poètes. C'est dans les glorieuses colonnes du fait positif et des mesures légales qu'il faut chercher la vraie beauté. Tenez, Mrs Sampson, dis-je, il y a des chiffres, qui surpassent en merveilles tous les poèmes, jusque dans ce tronc d'arbre sur lequel nous sommes assis. Les cercles de l'aubier montrent qu'il est mort à 60 ans. Enterré à 600 mètres de profondeur, il se transformerait en charbon d'ici 3 000 ans. La mine de charbon la plus profonde de la Terre se situe à Killingworth, près de Newcastle. Une caisse longue de 4 pieds, large de 3 pieds et haute de 2 pieds 8 pouces peut contenir une tonne de charbon. Lorsqu'une artère est rompue, comprimez-la au-dessus de la blessure. Une jambe humaine possède trente os. La tour de Londres fut brûlée en 1841.

Continuez, Mr. Pratt, dit Mrs Sampson. De telles pensées sont si originales et revigorantes! J'estime que les statistiques sont aussi admirables qu'il est possible de l'être.» (*Le Manuel du Mariage*)

Je trouve que ce passage est très beau. Et, naturellement, l'ouvrage est un bon exemple de clarté dans un texte savant. Mais je me demande quel rapport il peut y avoir entre le passage cité et le contenu du livre. Les auteurs y voient-ils un lien étroit? le passage, qui semble faire allusion au genre de préoccupation que peut avoir un praticien de la statistique, est-il cité pour justifier ou expliquer la superstructure de la statistique probabiliste qui se bâtit sur le travail des praticiens?

Les auteurs pensent-ils que les concepts et les formules développés dans leur traitement de la statistique doivent guider ou fonder les travaux des praticiens et des bureaux de la statistique? Ou bien pensent-ils que la justification de la statistique probabiliste réside dans le fait que ses techniques sont utilisées dans les productions des bureaux de la statistique?

Qu'y a-t-il de *réel* et d'*imaginatoire* dans ce rapport?

Il y a un peu d'énigme dans le rapport qui peut exister entre les travaux des théoriciens et des praticiens. Les choses semblent se passer comme ceci :

- les statistiques produites ont une valeur parce que leur production repose sur des méthodes reconnues;
- la méthode elle-même découle d'une théorie;
- mais la théorie statistique est formulée au moyen de concepts et de la logique mathématique, eux-mêmes fondés sur des postulats invérifiables.

Qu'est-ce qui justifie les postulats, les concepts et la théorie?

En science, de façon générale, une théorie est jugée valable par l'utilité des résultats que permet d'obtenir la technique qui en découle.

En réalité, une technique est souvent inventée sans *théorie* tandis que son utilité en fait répandre l'emploi. Le bronze et l'acier damassé ont été mis au point parce qu'ils étaient utiles, non parce qu'on avait formulé une théorie mathématique cohérente à propos de la métallurgie.

Voyons maintenant des exemples de ce qui se fait comme travaux dans nos deux catégories de l'activité statistique. Le journal officiel de la Société statistique du Canada est la *Revue Canadienne de Statistique*, qui est une publication trimestrielle. L'organe officiel de Statistique Canada est *Le Quotidien*, qui a paru 256 fois l'an dernier. Une comparaison des titres des articles publiés offre un grand intérêt. J'ai choisi au hasard dans *La Revue Canadienne de Statistique* 15 expressions clés parmi 122 qui pouvaient être représentatives des articles publiés en 1983. Cette liste d'expressions donne une idée des sujets sur lesquels les théoriciens de la statistique lisent et écrivent.

- Distributions d'abondance
 - Propriétés asymptotiques
 - Distribution centrée de Wishart
 - Distribution de khi carré
 - Valeurs limites d'acceptation
 - Théorie de la décision
 - Analyse de courbe de croissance
 - Filtre linéaire
 - Processus logarithmique
 - Etudes longitudinales
 - Modèle linéaire à plusieurs variables
 - Estimateur de mouvement
 - Séries chronologiques spatiales
 - Propriétés de structure
 - Estimation (pondérée) par les moindres carrés
- Ces expressions sont dans toutes les bouches à cette conférence. Mais elles désignent des thèmes qui *ne sont pas* du tout ceux dont s'occupent les praticiens de la statistique. D'ailleurs beaucoup de praticiens, la plupart peut-être, ne connaissent pas ces expressions ou ne s'y intéressent.

Les communiqués parus dans *Le Quotidien* du 29 avril 1925 donnent une idée de la production de Statistique Canada.

- nombre total de porcs au Canada (plus de dix millions)
- nombre de tonnes d'orge exportées (plus de 150,000 en mars 1985)
- nombre de mètres carrés de laine minérale expédiés (plus de six millions)

On peut encore se faire une idée de la production de Statistique Canada au moyen du tableau des principaux indicateurs statistiques mis à jour toutes les semaines dans la publication faits saillants, distribuée aux ministres et sous-ministres. Voici quelques-uns de ces indicateurs :

- Produit national brut
- Mises en chantier de logements
- Taux d'escompte
- Taux de chômage
- Hausse de l'indice des prix à la consommation
- Gains hebdomadaires

Il y a aussi les mesures qui ont trait à d'autres secteurs de la société canadienne comme l'économie, les affaires, le commerce, les finances, les services sociaux et le monde du travail. Statistique Canada produit des études sur des sujets comme le divorce au Canada, la santé des canadiens, la condition féminine, les indicateurs de conjoncture, les indicateurs de la science et de la technologie, les caractéristiques linguistiques des canadiens et ainsi de suite. Je tiens à préciser que je ne cherche pas à porter un jugement sur la valeur relative des deux types de production statistique dont je parle ici. Ces deux genres d'activité sont utiles à la société, comme en témoigne le fait que l'une et l'autre a ses utilisateurs. Du point de vue social, ces deux types d'activité sont l'un et l'autre justifiés.

Statisticiens et statisticiens¹

MARTIN B. WILK²

Je suis honoré d'avoir été invité à prononcer une allocution à cette conférence annuelle de la Société statistique du Canada.

Mais à cet honneur est attachée, je le crains, la responsabilité de dire des choses intéressantes. C'est une tâche qui n'est pas facile. C'est pourquoi j'ai pensé procéder par étapes, en choisissant d'abord un titre pour mon exposé. Puis en tâchant d'expliquer le sens de ce titre. Cette explication allait constituer le fond de mon allocution. Malheureusement, je ne suis pas encore certain de ce que veut dire le titre que j'ai choisi. Mais je ne me laisserai pas arrêter par cela. Évidemment, comme disait Yogi Berra, si vous ne savez pas où vous allez, il se peut que vous n'y arriviez jamais.

Il y a beaucoup de gens à qui l'on donne le titre de statisticien et qui font des travaux très divers dont on peut dire qu'ils sont de la statistique. En fait, à diverses époques de ma carrière variée, j'ai moi-même été plusieurs sortes de statisticien. Cette observation me conduit à me demander quelles relations peuvent exister entre les différents genres de statisticien et de statistique.

On pourrait définir deux types de statistique: d'abord la statistique probabiliste, puis la production de bases de données dont s'occupent les bureaux de la statistique. Que faut-il entendre par statistique probabiliste? Sans chercher à être précis, je dirai que cette expression englobe les matières normalement traitées dans les manuels et les cours, par exemple l'analyse de variance, les tests de validité de l'ajustement, les plans d'expériences, les composantes de la variance, l'estimation par la méthode de Bayes et ainsi de suite.

Quant aux résultats des bureaux de la statistique comme Statistique Canada, le Bureau de la statistique du Manitoba et le U.S. Bureau of the Census, vous les lisez tous les jours dans les journaux.

Ces deux genres de travaux sont *considérés* comme liés entre eux et, je crois, *le sont*. On pourrait dire que la relation entre les deux est à la fois *réelle* et *imaginaire* — et je ne suis pas du tout certain de ce qu'il convient de ranger dans l'une et l'autre catégorie. Considérons donc certaines réalités à quoi correspondent ces deux catégories, que nous appellerons la statistique théorique et la statistique pratique tout simplement pour ne pas employer des expressions comme "statisticien probabiliste".

La Société statistique du Canada semble composée surtout de théoriciens de la statistique. En effet, un sondage récent a montré qu'elle compte 66% d'universitaires et 21% de membres du personnel d'organismes gouvernementaux. Des 2,000 spécialistes que comprend la société, 32 seulement appartiennent à Statistique Canada.

Les personnes inscrites à cette conférence sont vraisemblablement pour la plupart des théoriciens qui donc s'intéressent avant tout à la statistique probabiliste. Non seulement n'y a-t-il que quelques personnes de Statistique Canada, mais j'ai dû aussi constater que les chefs de service de Statistique Canada ne se sont pas montrés très intéressés à envoyer de leur personnel à la conférence.

¹ Allocution prononcée à la conférence annuelle de la Société statistique du Canada.
² M. Martin B. Wilk, ancien statisticien en chef du Canada, est actuellement conseiller principal auprès du Conseil privé et président de la Société statistique du Canada.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

Volume 11, numéro 2, décembre 1985

TABLE DES MATIÈRES

A.B. WILK	Statisticiens et statisticiens	101
D. DREW, Y. BÉLANGER, et P. FOY	La stratification dans l'enquête sur la population active du Canada	109
D.R. BELLHOUSE	Échantillonnage des questionnaires manuscrits de recensement reproduits sur microfilm	125
B.C. SAXENA, P. NARAIN, et A.K. SRIVASTANA	Estimation du total pour deux caractères dans les enquêtes à bases de sondage multiples	135
B.B. DAGUM et M. MORRY	Désaisonnalisation des séries pour la population active en périodes de récession et de non-récession	149
B.B. DAGUM, G. HUOT, N. GAIT, et N. LANIEL	Formes de relation entre le niveau total de chômage et le nombre de bénéficiaires de l'assurance-chômage au Canada	163
L. SWAIN	Principes fondamentaux pour le développement des questionnaires	181
R.B.P. VERMA et P. PARENT	Vue d'ensemble des avantages et inconvénients des fichiers de données administratives choisis	193
R.D. SHARMA et C. WONG	Utilisation de fichiers de données administratives pour la production d'estimations de la migration: une étude de cas du fichier des permis de conduire en Ontario	203
F. AHMAD, R. CHOW, O. DEVRIES, A. HASHMI, et Y. MARCOGLIESE	Utilisation des dossiers de l'assurance-maladie de l'Alberta pour la production d'estimations régionales de la population	211
D.G. McRAE	Utilisation des comptes de l'Hydro dans le modèle d'estimation par régression de la population en Colombie-Britannique	223
D.S. O'NEIL et C.D. MCINTOSH	Estimation de la répartition par âge et par sexe de la population totale des petites régions	229
R.B.P. VERMA, K.G. BASAVARAJAPPA, et R.K. BENDER	Estimation de la population par âge et par sexe	237
R.K. BENDER	Estimation de la population des petites régions: expérience de Statistique Canada	247
	Collaboration à la rédaction	252

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

COMITÉ DE RÉDACTION

Président

R. Platek, *Statistique Canada*

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*

D.R. Bellhouse, *Université Western Ontario*

E.B. Dagum, *Statistique Canada*

J.F. Gentleman, *Statistique Canada*

G.J.C. Hole, *Statistique Canada*

T.M. Jeays, *Statistique Canada*

G. Kalton, *Université du Michigan*

C. Patrick, *Statistique Canada*

J.N.K. Rao, *Université Carleton*

C.E. Särndal, *Université de Montréal*

V. Tremblay, *Université de Montréal*

H. Lee, *Statistique Canada*

COMITÉ DE DIRECTION

R. Platek (Président), E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

POLITIQUE DE RÉDACTION

La revue *Techniques d'enquête* publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les techniques de lissage et d'extrapolation, les études démographiques, l'intégration et l'analyse de production de statistiques. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles sont soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de *Statistique Canada*.

Présentation de textes pour la revue

La revue *Techniques d'enquête* est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociale, *Statistique Canada*, 4^e étage, Edifice Jean-Talon, Tunney's Pasture, Ottawa (Ontario), Canada KIA 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue *Techniques d'enquête* (catalogue n° 12-001) est de 10,00\$ par copie, 20,00\$ par année au Canada, et de 11,50\$ par copie, 23,00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Vente et distribution des publications, *Statistique Canada*, Ottawa (Ontario), Canada KIA 0T6.

Statistique Canada

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

Décembre 1985

Publication autorisée par
le ministre des Approvisionnements
et Services Canada

© Ministre des Approvisionnements
et Services Canada 1986

Mai 1986

8-3200-501

Prix: Canada, \$10.00, \$20.00 par année
Autres pays, \$11.50, \$23.00 par année

Paiement en dollars canadiens ou l'équivalent
Catalogue 12-001, vol. 11, n° 2

ISSN 0714-0045

Ottawa

TECHNIQUES D'ENQUÊTE

Statistique Canada Statistics Canada



UNE REVUE
DE
STATISTIQUE CANADA

VOLUME 11, NUMÉRO 2
DÉCEMBRE 1985

Canada

001



Statistics Canada · Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA



VOLUME 12, NUMBER 1
JUNE 1986

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

JUNE 1986

Published under the authority of
the Minister of Supply and
Services Canada

©Minister of Supply
and Services Canada 1986

November 1986
8-3200-501

Price: Canada, \$10.00, \$20.00 a year
Other Countries, \$11.50, \$23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 12, No. 1

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

EDITORIAL BOARD

Chairman R. Platek, *Statistics Canada*

Editor M.P. Singh, *Statistics Canada*

Associate Editors

K.G. Basavarajappa, *Statistics Canada*

D.R. Bellhouse, *University of Western Ontario*

L. Biggeri, *University of Florence*

E.B. Dagum, *Statistics Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistics Canada*

G.J.C. Hole, *Statistics Canada*

T.M. Jeays, *Statistics Canada*

G. Kalton, *University of Michigan*

C. Patrick, *Statistics Canada*

J.N.K. Rao, *Carleton University*

C.E. Särndal, *University of Montreal*

F.J. Scheuren, *U.S. Internal Revenue Service*

V. Tremblay, *Statplus, Montreal*

K.M. Wolter, *U.S. Bureau of the Census*

Assistant Editors

J. Armstrong, *Statistics Canada*

H. Lee, *Statistics Canada*

MANAGEMENT BOARD

R. Platek (Chairman), J. Armstrong, E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

EDITORIAL POLICY

The Survey Methodology Journal will publish articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis will be on the development and evaluation of specific methodologies as applied to actual data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$10.00 per copy, \$20.00 per year in Canada, \$11.50 per copy, \$23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. (A reduced price is available to members of some statistical organizations. Please check with and subscribe through your organization.)

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 12, Number 1, June 1986

Special Edition – Missing Data in Surveys

CONTENTS

Preface	0
G. KALTON and D. KASPRZYK The Treatment of Missing Survey Data	1
R. PLATEK and G.B. GRAY On the Definitions of Response Rates	17
V.P. GODAMBE and M.E. THOMPSON Some Optimality Results in the Presence of Nonresponse	29
D.B. RUBIN Basic Ideas of Multiple Imputation for Nonresponse	37
P. GILES and C. PATRICK Imputation Options in a Generalized Edit and Imputation System	49
M.S. SRIVASTAVA and E.M. CARTER The Maximum Likelihood Method for Nonresponse in Sample Surveys	61
M.A. HIDIROGLOU and J.-M. BERTHELOT Statistical Editing and Imputation for Periodic Business Surveys	73
V. TREMBLAY Practical Criteria for Definition of Weighting Classes	85
S. CHEUNG and C. SEKO A Study of the Effects of Imputation Groups in the Nearest Neighbour Imputation Method for the National Farm Survey	99

PREFACE

This issue is devoted to papers presented at the Methodology Symposium on Missing Data in Surveys held at Statistics Canada in Ottawa, April 16-17, 1986. The symposium was jointly sponsored by Statistics Canada's Methodology Research Committee and the Laboratory for Research in Statistics and Probability at Carleton University. Concern about missing data in surveys (due to non-response or unusable responses) has been increasing in recent years. The symposium provided a forum for more than 200 professionals from universities, government organizations and the private sector in Canada and the United States to exchange information concerning recent theoretical and applied developments.

The symposium was opened by the Chief Statistician of Canada, Dr. Ivan Fellegi. He spoke about the international community's concern about the growing gap between theoretical and applied statistics and commended the organizers for bringing together specialists from both fields. While stating that the primary purpose of the conference was to make headway in the chosen topic, Dr. Fellegi also noted that the underlying theme was the extent to which statistical agencies should be involved in model-building.

The symposium included four sessions. The first session "General Issues and Organizational Experiences" was chaired by L. Kish of the University of Michigan and included presentations by G. Kalton (University of Michigan), G.B. Gray (Statistics Canada), D.W. Chapman (U.S. Bureau of the Census) and L.R. Curtin (U.S. National Center for Health Statistics). The chairman of the afternoon session of April 16, "Design and Estimation" was M. Hansen of Westat Inc. Papers were presented by P.S.R.S. Rao (University of Rochester), S. Michaud (Statistics Canada), C.E. Särndal (University of Montreal), G. Lazarus (Statistics Canada) and V.P. Godambe (University of Waterloo).

The morning session of April 17, "Item Non-Response and Imputation" was chaired by M. Moore of the University of Montreal. This session included contributions by D. Rubin (Harvard University), P. Giles (Statistics Canada), M.S. Srivastava (University of Toronto) and M.A. Hidirolou (Statistics Canada). The chairman of the final session, "Case Studies", was J.N.K. Rao of Carleton University. Papers were presented by S. Hinkins (U.S. Internal Revenue Service), V. Tremblay (University of Montreal) and S. Cheung (Statistics Canada). The symposium was closed with a general discussion of developments concerning missing data in surveys led by J.N.K. Rao (chairman) and a panel including G. Kalton, L. Kish, D. Rubin, and I. Sande (Statistics Canada).

Nine of the symposium papers are included in this issue of the Journal. Additional symposium papers accepted for publication will appear in the next issue.

The Treatment of Missing Survey Data

GRAHAM KALTON and DANIEL KASPRZYK¹

ABSTRACT

Missing survey data occur because of total nonresponse and item nonresponse. The standard way to attempt to compensate for total nonresponse is by some form of weighting adjustment, whereas item nonresponses are handled by some form of imputation. This paper reviews methods of weighting adjustment and imputation and discusses their properties.

KEY WORDS: Nonresponse; Item nonresponse; Weighting adjustments; Imputation.

1. INTRODUCTION

Surveys typically collect responses to a large number of items for each sampled element. The problem of missing data occurs when some or all of the responses are not collected for a sampled element or when some responses are deleted because they fail to satisfy edit constraints. It is common practice to distinguish between total (or unit) nonresponse, when none of the survey responses are available for a sampled element, and item nonresponse, when some but not all of the responses are available. Total nonresponse arises because of refusals, inability to participate, not-at-homes, and untraced elements. Item nonresponse arises because of item refusals, "don't knows", omissions and answers deleted in editing.

This paper reviews the general-purpose methods available for handling missing survey data. The distinction between total and item nonresponse is useful here since different adjustment methods are used for these two cases. In general the only information available about total nonrespondents is that on the sampling frame from which the sample was selected (e.g., the strata and PSUs in which they are located). The important aspects of this information can usually be readily incorporated into weighting adjustments that attempt to compensate for the missing data. Hence as a rule weighting adjustments are used for total nonresponse. Methods for making weighting adjustments are reviewed in Section 2.

In the case of item nonresponse, however, a great deal of additional information is available for the elements involved: not only the information from the sampling frame, but also their responses for other survey items. In order to retain all survey responses for elements with some item nonresponses, the usual adjustment procedure produces analysis records that incorporate the actual responses to items for which the answers were acceptable and imputed responses for other items. Imputation methods for assigning answers for missing responses are reviewed in Section 3.

In general the choice between weighting adjustments and imputation for handling missing survey data is fairly clearcut; there are cases, however, when the choice is not so clear. These are cases of what may be termed partial nonresponse, when some data are collected for a sampled element but a substantial amount of data is missing. Partial nonresponse can arise, for instance, when a respondent terminates an interview prematurely, when data are not obtained for one or more members of an otherwise cooperating household (for household level analysis), or when a sampled individual provides data for some but not all waves of a panel survey. Discussions of the choice between weighting and imputation to compensate for wave nonresponse in a panel survey are given by Cox and Cohen (1985) and Kalton (1986).

¹ Graham Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48106-1248 and Daniel Kasprzyk, Population Division, U.S. Bureau of the Census, Washington, D.C., 20233. The authors would like to thank the referees for their helpful comments.

Although weighting adjustments and imputation are treated as separate approaches in the discussion below, they are in fact closely related. The relationship and differences between the two approaches are briefly discussed in Section 4, which also mentions some alternative ways of handling missing survey data.

2. WEIGHTING ADJUSTMENTS

Weighting adjustments are primarily used to compensate for total nonresponse. The essence of all weighting adjustment procedures is to increase the weights of specified respondents so that they represent the nonrespondents. The procedures require auxiliary information on either the nonrespondents or the total population. The following four types of weighting adjustments are briefly reviewed below: population weighting adjustments, sample weighting adjustments, raking ratio adjustments, and weights based on response probabilities. More details are provided in Kalton (1983).

2.1 Population Weighting Adjustments

The auxiliary information used in making population weighting adjustments is the distribution of the population over one or more variables, such as the population distribution by age, sex and race available from standard population estimates. The sample of respondents is divided into a set of classes, termed here weighting classes, defined by the available auxiliary information (e.g., White males aged 15-24, non-White females aged 25-34, etc.). The weights of all respondents within a weighting class are then adjusted by the same multiplying factor, with different factors in different classes. The adjustment is carried out in such a way that the weighted respondent distribution across the weighting classes conforms to the population distribution.

This type of adjustment is often termed poststratification. That term is avoided here, however, because although population weighting resembles poststratification, there is an important difference between the two. Like population weighting, poststratification weights the sample to make the sample distribution conform to the population distribution across a set of classes (or strata). However, the standard textbook theory of poststratification is concerned only with the sampling fluctuations that cause the sample distribution to deviate from the population distribution, not with the more major deviations that can arise from varying response rates across the classes. Poststratification adjustments are more like a fine tuning of the sample, resulting generally in only small variations in the weights across strata. In consequence, provided that the strata are not small, poststratification leads to lower standard errors for the survey estimates. In contrast, population weighting adjustments may involve more major adjustments and result in higher standard errors.

Population weighting adjustments attempt to reduce the bias created by nonresponse and coverage errors. Consider the estimation of a population mean \bar{Y} from a sample in which the elements are selected with equal probability. Suppose that the population is divided into a set of weighting classes, with a proportion W_h of elements in class h . Assume that respondents always respond and that nonrespondents never do. Let R_h and M_h be the proportions of respondents and nonrespondents respectively in class h , and let $\bar{R} = \sum W_h R_h$ be the overall response rate. Then, following Thomsen (1973), the bias of the unadjusted respondent mean (\bar{y}) can be expressed as

$$B(\bar{y}) = \bar{R}^{-1} \sum W_h (\bar{Y}_{rh} - \bar{Y}_r) (R_h - \bar{R}) + \sum W_h M_h (\bar{Y}_{rh} - \bar{Y}_{mh}) = A + B \quad (1)$$

where \bar{Y}_{rh} and \bar{Y}_{mh} are the means for respondents and nonrespondents in class h respectively, and \bar{Y}_r is the population mean for the respondents. The use of the population weighting adjustment leads to the weighted sample mean, $\bar{y}_p = \sum W_h \bar{y}_{rh}$, where \bar{y}_{rh} is the respondent sample mean in class h . The bias of \bar{y}_p is simply the second term in $B(\bar{y})$, that is, $B(\bar{y}_p) = B$.

If A and B are of the same sign, the population weighting adjustment reduces the absolute bias in the estimate of \bar{Y} by $|A|$. If $\bar{Y}_{rh} = \bar{Y}_{mh}$, as occurs in expectation when the nonrespondents are missing at random within the weighting classes, then $B = 0$. In this case, the population weighting adjustment eliminates the bias. The term A is a covariance-type term between the class response rates and the class respondent means. It is zero if either the response rates or the respondent means do not vary between classes. In either of these cases, the population weighting adjustment has no effect on the bias of the estimator. It may be noted that population weighting adjustments may increase the absolute bias of the estimate of \bar{Y} . This will occur when A and B are of opposite signs and $|A| < 2|B|$.

Population weighting adjustments require external data on the population distributions for the variables to be used. Care is needed to ensure that the data on which the population distributions are based are exactly comparable with the survey data; otherwise, inappropriate weights will result. Since the procedure weights up to population distributions, it does more than just attempt to compensate for nonresponse. It also compensates for coverage errors and makes a poststratification adjustment.

2.2 Sample Weighting Adjustments

As with population weighting adjustments, with sample weighting adjustments the sample is divided into weighting classes; varying weights are then assigned to these classes in an attempt to reduce the nonresponse bias. The essential difference between the two procedures lies in the auxiliary information used. As described above, population weighting adjustments are based on externally obtained population distributions. No data are needed for the sample nonrespondents. In contrast, sample weighting adjustments employ only data internal to the sample and require information about the nonrespondents.

With sample weighting adjustments, the nonresponse adjustment weights for the weighting classes are made proportional to the inverses of the response rates in the classes. In order to compute these response rates, the numbers of respondents and nonrespondents in the classes must be determined. It is therefore necessary to know to which class each respondent and nonrespondent belongs. Since typically very little information about the nonrespondents is available, the choice of weighting class is usually severely restricted. It is often limited to general sample design variables (e.g., PSUs and strata), characteristics of those variables (e.g., urban/rural, geographical region), and sometimes some additional variables available on the sampling frame. On occasion it may also be possible to collect information on one or two variables for the nonrespondents, for instance by interviewer observation.

As population weighting adjustments resemble poststratification, so sample weighting adjustments resemble two-phase sampling. The first phase sample is the total sample of respondents and nonrespondents; the second phase sample is the subsample of respondents, selected with different sampling fractions (response rates) in different strata (weighting classes). The sample weighted mean can be represented by $\bar{y}_s = \sum w_h \bar{y}_{rh}$, where w_h is the proportion of the total sample in weighting class h . Assuming no coverage errors, $E(w_h) \doteq W_h$, the population proportion in class h , as used in the population weighted estimator

$\bar{y}_p = \sum W_h \bar{y}_{rh}$. The bias of \bar{y}_s is the same as that of \bar{y}_p , namely $B(\bar{y}_s) = B$ as given in equation (1); hence the effect of the sample weighting adjustment on the bias of the survey estimate is the same as that of the population weighting adjustment. Since sample weighting adjustments use only data for the sample, they do not compensate for coverage errors (unlike population weighting adjustments).

Population and sample weighting adjustments have different data requirements, and hence address different potential sources of bias. In practice the two forms of adjustment are used in combination. Generally sample weighting adjustments are applied first, and then population weighting adjustments are applied afterwards. A common approach is initially to determine the sample weights needed to compensate for unequal selection probabilities, next to revise these weights to compensate for unequal response rates in different sample weighting classes (e.g., urban/rural classes within geographical regions), and finally to revise the weights again to make the weighted sample distribution for certain characteristics (e.g., age/sex) conform to the known population distribution for those characteristics. The use of this approach in the U.S. Current Population Survey is described by Bailer *et al.* (1978).

As with population weighting adjustments, the aim of sample weighting adjustments is to reduce the bias that nonresponse may cause in survey estimates. An effect of sample weighting adjustments is, however, to increase the variances of the survey estimates. There is therefore a trade-off to be made between bias reduction and variance increase.

An indication of the amount of increase in variance from weighting can be obtained by considering the situation where the element variances within the weighting classes are all the same and the variances between the class means are negligible compared to the within-class variances. In this situation, the loss of precision from weighting is approximately the same as that arising from the use of disproportionate stratified sampling when proportionate stratified sampling is optimum; Kish (1965, Section 11.7C; 1976) discusses this latter case.

Under the above conditions, weighting increases the variance of a sample mean by approximately $L = (\sum W_h k_h) (\sum W_h / k_h)$, where W_h is the proportion of the population and k_h is the weight for class h . An alternative expression for L is $(\sum n_h) (\sum n_h k_h^2) / (\sum n_h k_h)^2$, where n_h is the sample size in class h . The factor L becomes large when the variance of the weights is large.

A large variance in the weights can arise from segmenting the sample into many weighting classes with only a few sampled elements in each. When the weighting classes are small, their response rates are unstable, and this gives rise to a large variation in the weights. To avoid this effect, it is common practice to limit the extent to which the sample is segmented. Even so, there may still be some weighting classes that require large weights. Sometimes these weighting classes are handled by collapsing them with adjacent ones and sometimes their weights are cut back to some acceptable maximum value (see Bailer *et al.* 1978 and Chapman *et al.* 1986, for examples). These procedures avoid the increase in variance associated with the use of extreme weights, but they may lead to increased bias; their effect on the bias is, however, unknown.

In some cases it seems desirable to use several auxiliary variables in forming the weighting classes for population or sample weighting adjustments. However, if the classes are formed by taking the full crossclassification of the variables, there will be a large number of weighting classes. Unless the sample is very large, the sample sizes in the resultant weighting classes will be small, and the instability in the response rates will lead to a large variance in the weights and loss of precision in the survey estimates. One way to deal with this problem is to cut down on the number of classes by collapsing cells, for instance by discarding some of the auxiliary variables or using coarser classifications. Another way is to base the weights on a model, as is done in raking ratio weighting discussed below.

2.3 Raking Ratio Adjustments

When weighting classes are taken to be the cells in the crossclassification of the auxiliary variables, population weighting adjustments make the joint distribution of the auxiliary variables in the sample conform to that in the population. Similarly, sample weighting adjustments make the joint distribution of the auxiliary variables in the respondent sample conform to that in the total sample. As noted above, however, this crossclassification approach may have the undesirable effect of creating many small, and hence unstable, weighting classes. Also, it is not always possible to employ this approach with population weighting adjustments: in many cases the population marginal distributions, and perhaps some bivariate distributions, of the auxiliary variables are available, but the full joint distribution is unknown.

An alternative approach is to develop weights that make the marginal distributions of the auxiliary variables in the sample conform to marginal population distributions (with population weighting) or marginal total sample distributions (with sample weighting), without ensuring that the full joint distribution conforms. The method of raking ratio estimation, or raking, may be used to obtain weights that satisfy these conditions. Raking corresponds to iterative proportional fitting in contingency table analysis (see, for instance, Bishop *et al.*, 1975).

Consider the use of raking in the simple case of two auxiliary variables. Let W_{hk} be the proportion of the population in the (h, k) -th cell of the crossclassification, and let \tilde{w}_{hk} be the proportion assigned to that cell by the raking algorithm. Conditional on the total and respondent sample sizes in the cells (and assuming all cells have at least one respondent), the bias of the raking ratio adjusted sample mean $\bar{y}_q = \Sigma \Sigma \tilde{w}_{hk} \bar{y}_{hk}$ is

$$B(\bar{y}_q) = \Sigma \Sigma W_{hk} M_{hk} (\bar{Y}_{rhk} - \bar{Y}_{mhk}) + \Sigma \Sigma (\tilde{W}_{hk} - W_{hk}) (\bar{Y}_{rhk} - \bar{Y}_{rh.} - \bar{Y}_{r.k} + \bar{Y}_r)$$

where $\tilde{W}_{hk} = E(\tilde{w}_{hk})$. The first term in this bias corresponds to the bias term B in equation (1) for the population and sample weighting adjustments. It is zero in expectation if the cell nonrespondents are random subsets of the cell populations. The second term is zero if either $\tilde{W}_{hk} = W_{hk}$ or there is no interaction in the \bar{Y}_{rhk} for this classification.

Underlying the raking ratio weighting procedure is a logit model for the cell response rates. With the model $\ln[R_{hk}/(1 - R_{hk})] = \alpha_h + \beta_k$ for the response rates in a two-way classification, $\tilde{W}_{hk} = W_{hk}$. Thus, under this model, the second term in $B(\bar{y}_q)$ is zero.

Further discussion of raking ratio weighting is given by Oh and Scheuren (1978a, 1978b, 1983). Oh and Scheuren (1978a) also provide a bibliography on raking.

2.4 Weighting with Response Probabilities

Although a number of methods for weighting with response probabilities have been proposed, this approach has not been widely adopted as an adjustment procedure. The basis of the approach is to assume that all population elements have probabilities (usually required to be non-zero) of responding to the survey. Some method is used to estimate the response probabilities for responding elements. These elements are then given nonresponse adjustment weights that are in inverse proportion to their estimated response probabilities.

An early application of this approach is the well-known procedure of Politz and Simmons (1949, 1950). A single (evening) call is made to each selected household, and during the course of the interview respondents are asked on how many of the previous five evenings they were at home at about the same time. Their response probabilities are then taken to be the fraction of the six evenings (including the one of the interview) that they were at home, and the inverses of these probabilities are used in the analysis. Note that the procedure does not deal with those who were out on all six evenings and those who refused.

Another approach for estimating response probabilities is to regress response status (1 for respondents, 0 for nonrespondents) on a set of variables available for both respondents and nonrespondents, using a logistic or probit regression. The predicted values from the regression for the respondents are then taken to be their response probabilities, and weights in inverse proportion to these predicted values are used in the analysis. A special case is when the predictor variables are dummy variables that identify a set of classes. The predicted response probabilities are then the class response rates, and the method reduces to a sample weighting adjustment. The method is most appropriate for situations where a good deal of information is available for the nonrespondents, as for instance when the nonrespondents are losses after the first wave of a panel survey. Little and David (1983) discuss the application of the method for panel nonresponse. It should be noted that if the regression is highly predictive of response status, the resultant weights will vary markedly, leading to a substantial loss in the precision of the survey estimates.

Drew and Fuller (1980, 1981) describe an approach for estimating response probabilities from the number of respondents secured at successive calls. In their model, the population is divided into classes. Within each class, every element is assumed to have the same response probability which remains the same at each call. The model also allows for a proportion of hard-core nonrespondents that is assumed constant across classes. Under these assumptions, the response probabilities for each class and the proportion of hard-core nonrespondents can be estimated, and hence weighting adjustments can be made. Thomsen and Siring (1983) adopt a similar approach using a more complex model.

Finally, mention should be made of a related approach that compensates for nonresponse by weighting up difficult-to-interview respondents. Bartholomew (1961), for instance, proposed making only two calls in a survey, and weighting up the respondents at the second call to represent the nonrespondents. The assumption behind this approach is that the nonrespondents are like the late respondents. This assumption seems questionable, however, and empirical evidence from an intensive follow-up study of nonrespondents in the U.S. Current Population Survey does not support it (Palmer and Jones 1966; Palmer 1967).

3. IMPUTATION

A wide variety of imputation methods has been developed for assigning values for missing item responses. The aim here is to provide a brief overview of the methods, the basic differences between them, and some of the issues involved in imputation. A fuller treatment is provided by Kalton and Kasprzyk (1982).

Imputation methods can range from simple *ad hoc* procedures used to ensure complete records in data entry to sophisticated hot-deck and regression techniques. The following are some common imputation procedures:

- (a) *Deductive imputation.* Sometimes the missing answer to an item can be deduced with certainty from the pattern of responses to other items. Edit checks should check for consistency between responses to related items. When the edit checks constrain a missing response to only one possible value, deductive imputation can be employed. Deductive imputation is the ideal form of imputation.
- (b) *Overall mean imputation.* This method assigns the overall respondent mean to all missing responses.
- (c) *Class mean imputation.* The total sample is divided into classes according to values of the auxiliary variables being used for the imputation (comparable to weighting classes). Within each imputation class the respondent class mean is assigned to all missing responses.

- (d) *Random overall imputation.* A respondent is chosen at random from the total respondent sample, and the selected respondent's value is assigned to the nonrespondent. This method is the simplest form of hot-deck imputation, that is an imputation procedure in which the value assigned for a missing response is taken from a respondent to the current survey.
- (e) *Random imputation within classes.* In this hot-deck method, a respondent is chosen at random within an imputation class, and the selected respondent's value is assigned to the nonrespondent.
- (f) *Sequential hot-deck imputation.* The term sequential hot-deck imputation is used here to describe the procedure used with the labor force items in the U.S. Current Population Survey (Brooks and Bailer 1978). The procedure starts with a set of imputation classes. A single value for the item subject to imputation is assigned for each class (perhaps taken from a previous survey). The records in the survey's data file are then considered in turn. If a record has a response for the item in question, its response replaces the value stored for the imputation class in which it falls. If the record has a missing response, it is assigned the value stored for its imputation class.

The hot-deck method is similar to random imputation within classes. If the order of the records in the data file were random, the two methods would be equivalent, apart from the start-up process. The non-random order of the list generally acts to the benefit of the hot-deck method since it gives a closer match of donors and recipients provided that the file order creates positive autocorrelation. The benefit is, however, unlikely to be substantial.

The sequential hot-deck suffers the disadvantage that it may easily make multiple uses of donors, a feature that leads to a loss of precision in survey estimates. Multiple use of a donor occurs when, within an imputation class, a record with a missing response is followed by one or more other records with missing responses. The number of imputation classes that can be used with the method also has to be limited in order to ensure that donors are available within each class.

Useful discussions of the sequential hot-deck method are provided by Bailer *et al.* (1978), Bailer and Bailer (1978, 1983), Ford (1983), Oh and Scheuren (1980), Oh *et al.* (1980), and Sande (1983).

- (g) *Hierarchical hot-deck imputation.* The above disadvantages of the sequential hot-deck are avoided in the hierarchical hot-deck method, a form of hot-deck imputation developed for the items in the March Income Supplement of the Current Population Survey. The procedure sorts respondents and nonrespondents into a large number of imputation classes from a detailed categorization of a sizeable set of auxiliary variables. Nonrespondents are then matched with respondents on a hierarchical basis, in the sense that if a match cannot be made in the initial imputation class, classes are collapsed and the match is made at a lower level of detail. Coder (1978) and Welniak and Coder (1980) provide further details on the hierarchical hot-deck procedure.
- (h) *Regression imputation.* This method uses respondent data to regress the variable for which imputations are required on a set of auxiliary variables. The regression equation is then used to predict the values for the missing responses. The imputed value may either be the predicted value, or the predicted value plus some residual. There are several ways in which the residual may be obtained, as discussed later.
- (i) *Distance function matching.* This hot-deck method assigns a nonrespondent the value of the "nearest" respondent, where "nearest" is defined in terms of a distance function for the auxiliary variables. Various forms of distance function have been proposed (e.g., Sande 1979; Vacek and Ashikago 1980), and the function can be constructed to reduce the multiple use of donors by incorporating a penalty for each use (Colledge *et al.* 1978).

Although at first sight these may appear a diverse set of procedures, they can nearly all be fitted within a single unifying framework. The methods can all be described, at least approximately, as special cases of the general regression model

$$\hat{y}_{mi} = b_{ro} + \sum b_{rj} z_{mij} + \hat{e}_{mi} \quad (2)$$

where \hat{y}_{mi} is the imputed value for the i th record with a missing y value, z_{mij} are values reflecting the auxiliary variables for that record, b_{ro} and b_{rj} are the regression coefficients for the regression of y on x for the respondents, and \hat{e}_{mi} is a residual chosen according to a specified scheme for the particular imputation method.

Equation (2) represents the regression imputation method in an obvious way. If the \hat{e}_{mi} 's are set at zero, then the imputed value is the predicted value from the regression; otherwise a residual of some form may be added. The equation also represents class mean imputation by defining the z_j 's to be dummy variables that represent the classes, and setting $\hat{e}_{mi} = 0$. The regression equation then reduces to $\hat{y}_{mi} = \bar{y}_{rh}$, the class mean. Random imputation within classes is obtained by adding a residual to the class mean, where the residual is the deviation from the class mean for one of the respondents. Then $\hat{y}_{mi} = \bar{y}_{rh} + e_{rhk}$, where e_{rhk} is the deviation for respondent k in class h ; this reduces to $\hat{y}_{mi} = y_{rhk}$, the value for that respondent. The sequential and hierarchical hot-deck methods resemble the random within class method. The overall mean and random overall imputation methods are degenerate cases of the class mean and random within class methods that use no auxiliary information.

An important consideration in the choice of imputation method is the type of variable being imputed. All the above methods can be applied routinely with continuous variables, but some of them are not suitable for use with categorical or discrete variables (such as being a member of the labor force (1) or not (0), and the number of completed years of education). Overall mean, class mean, and regression imputations impute values like 0.7 for being a member of the labor force (i.e., a 70% chance) and 10.7 for the number of completed years of education. These values are not feasible for individual respondents, and rounding them to whole numbers leads to bias. For this reason, these imputation methods do not work well for categorical and discrete variables. A notable advantage of all hot-deck methods is that they always give feasible values since the values are taken from respondents.

There are two major distinguishing features of the above imputation methods that deserve elaboration: whether or not a residual is added and, if one is, the form of the residual; and whether the auxiliary information is used in dummy variable form to represent classes or whether it is used straightforwardly in the regression. These features are discussed in the next two subsections. Other issues arising with the use of imputation are then discussed in subsequent subsections.

3.1 Choice of Residuals

Imputation methods may be classified as deterministic or stochastic according to whether the \hat{e}_{mi} 's are set at zero or not. For each deterministic imputation method, there is a stochastic counterpart. Let \hat{y}_{mid} be the value imputed by the deterministic method and $\hat{y}_{mis} = \hat{y}_{mid} + \hat{e}_{mi}$ be that imputed by the corresponding stochastic method. Then $E_2(\hat{y}_{mis}) = \hat{y}_{mid}$, where E_2 denotes expectation over the sampling of residuals given the initial sample, provided that $E_2(\hat{e}_{mi}) = 0$ (as generally applies).

The choice between a deterministic and the corresponding stochastic imputation method depends on the form of survey analysis to be conducted. Consider first the estimation of the population mean of the y -variable using the sample mean of the respondents' values and

the nonrespondents' imputed values. As Kalton and Kasprzyk (1982) show, given that $E_2(\hat{y}_{mis}) = \hat{y}_{mid}$, it follows that the expectation of the sample mean is the same whether the deterministic method or the corresponding stochastic method is used. Thus both methods have the same effect on the bias of the estimate. However, the addition of random residuals in the stochastic method causes a loss of precision in the sample mean. Although this loss can be controlled by the choice of a suitable method of sampling residuals (Kalton and Kish 1984), nevertheless some loss in precision occurs. For this reason a deterministic scheme is preferable for the purpose of estimating the population mean.

Consider now the estimation of the element standard deviation and distribution of the y -variable. Deterministic imputation methods fare badly for these purposes, since they cause an attenuation in the standard deviation and they distort the shape of the distribution. This may be simply illustrated in terms of the class mean imputation method. By assigning the class mean to all the missing values in a class, the shape of the distribution is clearly distorted with a series of spikes at the class means. The standard deviation of the distribution is attenuated because the imputed values reflect only the between-class and not the within-class variance. The appeal of the stochastic imputation methods is that the residual term captures the within-class (or residual) variance, and hence avoids the attenuation of the element standard deviation and the distortion of the distribution.

Since some survey analyses are likely to involve the distributions of the variables, stochastic imputation methods like the hot-deck methods are generally preferred. Once a decision is made to use a stochastic method, the question of how to choose the residuals arises. If the standard regression assumptions are accepted, the residuals could be chosen from a normal distribution with a mean of zero and a variance equal to the residual variance from the respondent regression. However, this places complete reliance on the model. An alternative that avoids the normality assumption is to choose the residuals randomly from the empirical distribution of the respondents' residuals. Another alternative is to select a residual from a respondent who is a "close" match to the nonrespondent, measuring "close" in terms of similar values on the auxiliary variables. This attractive alternative avoids the assumption of homoscedasticity and guards against misspecification of the distribution of the residual term. In the limit, the closest respondent is one who has the same values of all the auxiliary variables as the nonrespondent. In this case, the nonrespondent is given one of the matched respondents' values. This case arises with hot-deck methods, where nonrespondents and respondents are matched in terms of the auxiliary variables, and nonrespondents are assigned values from matched respondents.

A further consideration in the choice of residuals is to make the imputed values feasible ones. As noted above, deterministic methods may impute values for categorical and discrete variables that are not feasible. Some stochastic methods solve this problem through the allocation of the residuals. In particular, the use of respondents' residuals with the random within class and the sequential and hierarchical hot-deck methods ensures that the imputed values are feasible ones.

3.2 Imputation Class or Regression Imputation

As noted earlier, both imputation class and regression imputation methods fall within the imputation model given by equation (2). The difference between them lies in the ways in which they employ the auxiliary variables.

Imputation class methods divide the sample into a set of classes. For this purpose, continuous auxiliary variables have to be categorized. There is complete flexibility in the way the classes are formed, and the symmetrical use of the auxiliary variables in different parts

of the sample is not required. Thus, for instance, in imputing for hourly rate of pay in a sample of employees, the sample might first be divided into two parts, union members and nonmembers; then the imputation classes for the members might be formed in terms of age and occupation whereas those for nonmembers might be formed in terms of sex and industry. As a rule, the aim is to construct classes of adequate size that explain as much of the variance in the variable to be imputed as possible. When the classes are formed by a complete crossclassification of the auxiliary variables, the underlying model contains all main effects and all interactions for the crossclassification. The limitation of imputation class methods is that the number of classes formed has to be constructed to ensure that there is some minimum number of respondents in each class. The hierarchical hot-deck method attempts to extend the amount of auxiliary data used, but even with this method matches of respondents and nonrespondents often cannot be made at the finer levels of detail. Coupled with the use of a random respondent residual within a class, imputation class methods have the valuable property that imputed values are feasible ones: that is, the imputed values are actual respondents' values.

Regression imputation methods have an advantage over imputation class methods in the number and in the level of detail of the auxiliary variables they can employ. Age can, for instance, be taken as a continuous variable rather than being categorized into a few classes. The regression model allows more main effects to be included in the model, but at the price of fewer interactions. Regression models can, of course, include some interactions, but they need to be specified. The models can also include polynomial terms and employ transformations, but again they need to be specified. The regression model has the potential of providing better predictions for the imputed values, but to achieve this careful modelling is required. Careful imputation modelling is unrealistic for all the variables in a survey, but it may be feasible for one or two major ones (and especially so for continuous surveys). Without careful modelling, there is a serious risk of poor imputations, although as noted earlier, this risk can be reduced by the allocation of random residuals from "close" respondents.

If a regression imputation assigns the residual from a respondent with exactly the same values of the auxiliary variables, the imputed value is necessarily a feasible one. If, however, there is even a small difference between the respondent's and nonrespondent's values on the auxiliary variables, the imputed value may not be feasible. A variant of regression imputation that avoids this problem, termed predictive mean matching, is described by Little (1986b) (Little attributes the method to Rubin). With predictive mean matching, the nonrespondent is matched to the respondent with the closest predicted value. Then, instead of adding the respondent's residual to the nonrespondent's predicted value, the nonrespondent is assigned the respondent's value. The method is thus a hot-deck method, and is similar to distance function matching.

The choice between imputation class and regression imputation methods should in part depend on the efforts made to develop the regression model. Unless adequate resources are devoted to the development of a regression model, the imputation class methods may be safer. The choice should also in part depend on the sample size. With large samples, hot-deck methods are likely to be able to use enough classes to take advantage of all the major predictor variables; however, with small samples this may not hold, and regression methods may have greater potential. David *et al.* (1986) describe an interesting study that compares regression models for imputing wages and salary in the U.S. Current Population Survey with hierarchical hot-deck imputations. Despite the extensive efforts made to develop the regression models, the hot-deck imputations were not found to be inferior in this large sample.

3.3 Effect of Imputation on Relationships

Although most of the literature on imputation deals with its effect on univariate statistics such as means and distributions, a large part of survey analysis is concerned with bivariate

and multivariate relationships. Here the analysis of relationships can be considered in broad terms to include crosstabulation, correlation or regression analysis, comparisons of subclass means or proportions, and any other analysis involving two or more variables. As will be illustrated below, imputation can have harmful effects on all analyses of relationships, often attenuating the associations between variables. Discussions of the effects of imputations on relationships are provided by Santos (1981), Kalton and Kasprzyk (1982) and Little (1986a).

The general nature of the effect of imputation on relationships can be seen by considering its effect on the estimate of the sample covariance in the simple situation where the y -variable has missing responses that are missing at random over the population and the x -variable has no missing data. The sample covariance, s_{xy} , is calculated in the standard way, based on the actual values for respondents and the imputed values for nonrespondents, as an estimate of the population covariance S_{xy} . Using the fact that $E_2(\hat{y}_{mis}) = \hat{y}_{mid}$ as above, it can be readily shown that the expected value of s_{xy} under a deterministic imputation method is the same as that under the corresponding stochastic method.

As Santos (1981) shows, the relative bias of s_{xy} when the mean overall or random overall imputation methods are used is approximately $-\bar{M}$, where \bar{M} is the nonresponse rate. This occurs because the imputed y -values are unrelated to their x -values, and hence the cases with imputed values attenuate the covariance towards zero. This attenuation is decreased in magnitude by imputation methods that use auxiliary variables. With class mean imputation or random imputation within classes, the relative bias is approximately $-\bar{M}(S_{xy,z}/S_{xy})$, where $S_{xy,z} = \sum W_h S_{xyh}$ is the average within-class covariance for classes formed by the auxiliary variables z , S_{xyh} is the covariance within class h , and W_h is the proportion of the population in class h . With predicted regression imputation or regression imputation with a random residual, both with a single auxiliary variable z , the relative bias is approximately $-\bar{M}[1 - (\rho_{xz}\rho_{yz}/\rho_{xy})]$, where ρ_{uv} is the correlation between u and v .

The disturbing feature of these results is that, unless \bar{M} is small, s_{xy} calculated with imputed values under any of these imputation methods may be subject to substantial bias even under the missing at random model. The estimates s_{xy} computed with imputed values obtained under the imputation class and regression methods are unbiased only if the partial covariance $S_{xy,z}$ is zero. In general, there is no reason to assume uncritically that $S_{xy,z}$ is zero. However, there is an important case when $S_{xy,z} = 0$. This occurs when $x = z$, that is when x is used as an auxiliary variable in the imputation procedure. In this case, the sample covariance is unbiased under the missing at random model. This result suggests that if the relationship between x and y is to form an important part of the survey analysis, x should be used as an auxiliary variable in imputing for missing y -values.

The above theory assumes that only the y -variable was subject to missing data. In practice the x -variable will often also be incomplete. If so, the sample covariance may be attenuated because of the imputations for both variables. A special feature occurs when x and y are both missing for a record. If the two values are imputed separately, the covariance is attenuated, but if they are imputed jointly, using the same respondent as the donor of both values, the covariance structure is retained. This suggests that when a record has several missing related values, they should be taken from the same donor. Coder (1978) describes the use of joint imputation from the same donor in the March Income Supplement of the Current Population Survey.

As an illustration of how the above arguments about the attenuation of covariances apply to other forms of relationships, we will give a simple numerical example of the effect of imputation on the difference between two proportions. Let the variable of interest be whether an individual has a particular attribute or not, and suppose that one half of the respondents fail to answer this question. The missing responses are imputed by a random within class imputation method using two classes, A and B . The objective is now to compare the

Table 1
 Number of Respondents with the Attribute, and Number of
 Sampled Persons by Class, Sex and Response Status

	Class A			Class B		
	M	F	Total	M	F	Total
Respondents with the attribute	80	40	120	60	20	80
Total respondents	100	100	200	100	100	200
Nonrespondents	100	100	200	100	100	200
Total sample	200	200	400	200	200	400

percentages of men and women with the attribute. The data are displayed in Table 1. Since 60% of the total respondents in class *A* have the attribute, 60 of the 100 male and 60 of the 100 female nonrespondents in that class will be imputed to have the attribute. Similarly, in class *B* 40% of the total respondents have the attribute, and so 40 male and 40 female nonrespondents will be imputed to have the attribute. The proportion of actual and imputed males with the attribute is thus $(80 + 60 + 60 + 40)/400 = 0.6$ or 60%. For females the corresponding proportion is $(40 + 60 + 20 + 40)/400 = 0.4$, or 40%. The difference between these two percentages is 20%.

Had sex also been taken into account in forming the imputation classes, the percentages of males and females with the attribute would have been 70% and 30%, differing by 40%. The failure to include sex as an auxiliary variable in the imputation has thus caused a substantial attenuation in the measurement of the relationship between sex and having the attribute.

3.4 Multiple Imputations

Ideally the analyst using a data set with imputed values should be able to obtain valid results for any analyses by applying standard techniques for complete data. However, as noted in the last section, imputation can distort measures of the relationships between variables. It also distorts standard error estimation.

All imputation methods except deductive imputation fabricate data to some extent. The extent of fabrication depends on how well the imputation model predicts the missing values. If the imputation model explains only a small proportion of the variance in the variable among the respondents, the amount of fabrication in each imputed value is likely to be substantial. If the imputation model explains a high proportion of the respondent variance, the amount of fabrication is likely to be less serious. However, it needs to be recognized that the fit of the imputation model for the respondents is not necessarily a good measure of the fit for the nonrespondents.

Standard errors computed in the standard way from a data set with imputed values will generally be underestimates because of the fabrication involved in the imputed values. Rubin (1978, 1979) has advocated the method of multiple imputations to provide valid inferences from data sets with imputed values (see also Herzog and Rubin 1983; Rubin and Schenker 1986). When multiple imputations are used for the purpose of standard error estimation, the construction of the complete data set by imputing for the missing responses is carried out several (say m) times using the same imputation procedure. The sample estimates z_i ($i = 1, 2, \dots, m$) of the population parameter of interest Z are computed from each of the replicate data sets, and their average \bar{z} is calculated. A variance estimator for \bar{z} is then

given by $\hat{V} = \hat{W} + [(m + 1)/m]\hat{B}$, where \hat{W} is the average of the within-replicate variance of \bar{z} and $\hat{B} = \Sigma(z_i - \bar{z})^2 / (m - 1)$ is the between-replicate variance. Even with the inclusion of the between-replicate variance component, however, the coverages of confidence intervals for Z based on \hat{V} are still overstated, with the amount of overstatement increasing with the level of nonresponse.

This overstatement of the confidence levels can be addressed by modifying the imputation procedure, as described by Rubin and Schenker (1986). Their treatment considers the random overall imputation method, and one of their modifications allows for uncertainty about the population mean and variance in the following way. With the standard random overall imputation method, the conditional expected mean and variance of the imputed values are the sample respondents' mean and variance. With the modification, the expected mean and variance of the imputed values for a replicate are drawn at random from appropriate distributions. The imputed values are then a random selection of respondents' values, modified for the randomly-chosen mean and variance. When estimating the population mean, the effect of the changing expected mean and variance between replicates is to increase the between-replicate variance component in \hat{V} . This increase gives improved coverage for the resultant confidence intervals.

A major problem with the use of multiple imputations is the additional computer analysis needed, which increases as the number of replicates, m , increases. For this reason, a small value of m , such as $m = 2$, may be preferred. A small value of m may, however, result in a low level of precision for the variance estimator. Even with small m , it is questionable whether the multiple imputation approach is feasible for routine analyses. It may be best reserved for special studies, such as that described by Herzog and Rubin (1983).

In addition to providing appropriate standard errors, another advantage of multiple imputations from the same imputation procedure is that it reduces the loss of precision in survey estimates arising from the random selection of respondents to act as donors of imputed values (see Section 3.1). This loss is reduced with multiple imputations by averaging over the replicates. A small number of replicates serves well for this purpose. As noted earlier, Kalton and Kish (1984) describe alternative ways of selecting the sample of respondents to achieve this end.

A second major potential application of multiple imputations is to generate the imputations for the several replicates by different imputation procedures, making different assumptions about the nonrespondents. Suppose, for instance, that hourly rates of pay are to be imputed for some earners in the sample. One procedure that might be used is the random within class imputation method, which is based on an assumption that nonrespondents are missing at random within the classes. If it is thought that the nonrespondents might in fact come more heavily from those with higher rates of pay in each class, a simple modification to the random within class method might be to impute values that are, say, 50 cents above the donors' values. Other imputation procedures - for instance, using different imputation classes - could also be tried. Comparison of the survey estimates obtained from the data sets in which the different imputation procedures are applied then provides a valuable indication of the sensitivity of the estimates to the values imputed. If the estimates turn out to be very similar, they can be accepted with greater confidence; if they differ markedly, the estimates need to be treated with considerable caution.

4. CONCLUDING REMARKS

Weighting and imputation have been presented as two distinct methods for handling missing survey data, but in fact there is a close relationship between them. This may be illustrated

by considering any imputation method that assigns respondents' values to the nonrespondents. For univariate analyses, this process is equivalent to dropping the nonrespondents' records and adding the nonrespondents' weights to those of the donor respondents (Kalton 1986).

The differences between weighting and imputation emerge when one considers the multivariate nature of survey data. It is possible to impute for the responses of a total nonrespondent by taking all the responses from a single donor; however, weighting is generally simpler in this case and it avoids the loss of precision arising from the sampling of respondents to serve as donors. It is not practicable to use weighting to handle item nonresponse since it would result in different sets of weights for each item; this would cause serious difficulties for crosstabulations and other analyses of the relationships between variables.

Weighting is a single global adjustment that attempts to compensate for the missing responses to all the items simultaneously. Imputation, on the other hand, is item-specific. This difference has consequences for the way that the auxiliary data are used. In forming weighting classes, the focus is on determining classes that differ in their response rates. The choice of auxiliary variables to use in imputation, however, is primarily made in terms of their abilities to predict the missing responses.

An assumption underlying all the procedures reviewed in this paper is that once the auxiliary variables have been taken into account the missing values are missing at random. Thus, for instance, the nonrespondents are assumed to be like the respondents within weighting and imputation classes. This assumption can be avoided by using stochastic censoring models, as has been done by Greenlees *et al.* (1982) in imputing wages and salaries in the Current Population Survey. However, as Little (1986b) observes, these models are highly sensitive to the distributional assumptions made.

An alternative approach for handling missing survey data is to leave the values missing in the data set and let the analyst incorporate appropriate missing data models into the analysis (Little 1982). This approach has much to commend it, but the labor and computing time needed to implement it effectively preclude its use as a general purpose strategy. Rather, the approach seems best suited for a small range of special analyses. In order to permit the analyst to adopt this approach, it is essential that all imputed values be flagged to indicate they are not actual responses, so that they can then be dropped from the analysis.

Finally, we should note that all methods of handling missing survey data must depend upon untestable assumptions. If the assumptions are seriously in error, the analyses may give misleading conclusions. The only secure safeguard against serious nonresponse bias in survey estimates is to keep the amount of missing data small.

REFERENCES

- BAILAR III, J.C., and BAILAR, B.A. (1978). Comparison of two procedures for imputing missing survey values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 462-467.
- BAILAR, B.A., and BAILAR III, J.C. (1983). Comparison of the biases of the hot-deck imputation procedure with an "equal-weights" imputation procedure. In *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (Eds. W.G. Madow and I. Olkin), New York: Academic Press, 299-311.
- BAILAR, B.A., BAILEY, L., and CORBY, C.A. (1978). A comparison of some adjustment and weighting procedures for survey data. In *Survey Sampling and Measurement*, (Ed. N.K. Namboodiri), New York: Academic Press, 175-198.
- BARTHOLOMEW, D.J. (1961). A method of allowing for 'not at home' bias in sample surveys. *Applied Statistics*, 10, 52-59.

- BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analyses*. Cambridge, Mass: The MIT Press.
- BROOKS, C.A., and BAILAR, B.A. (1978). *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Policy Working Paper 3. U.S. Department of Commerce. Washington, D.C.: U.S. Government Printing Office.
- CHAPMAN, D.W., BAILEY, L., and KASPRZYK, D. (1986). Nonresponse adjustment procedures at the U.S. Census Bureau. *Survey Methodology*, forthcoming.
- CODER, J. (1978). Income data collection and processing from the March Income Supplement to the Current Population Survey. *The Survey of Income and Program Participation Proceedings of the Workshop on Data Processing*, February 23-24, 1978, (Ed. D. Kasprzyk), Chapter II. Washington, D.C.: U.S. Department of Health, Education and Welfare.
- COLLEDGE, M.J., JOHNSON, J.H., PARE, R., and SANDE, I.G. (1978). Large scale imputation of survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 431-436.
- COX, B.G., and COHEN, S.B. (1985). *Methodological Issues for Health Care Surveys*. New York: Marcel Dekker.
- DAVID, M., LITTLE, R.J.A., SAMUHEL, M.E., and TRIEST, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.
- DREW, J.H., and FULLER, W.A. (1980). Modelling nonresponse in surveys with callbacks. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-642.
- DREW, J.H., and FULLER, W.A. (1981). Nonresponse in complex multiphase surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 623-628.
- FORD, B.L. (1983). An overview of hot-deck procedures. In *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 185-207.
- GREENLEES, W.S., REECE, J.S., and ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- HERZOG, T.N., and RUBIN, D.B. (1983). Using multiple imputation to handle nonresponse in sample surveys. In *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 209-245.
- KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor: Survey Research Center, University of Michigan.
- KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, forthcoming.
- KALTON, G., and KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22-31.
- KALTON, G., and KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics - Theory and Methods*, 13(16), 1919-1939.
- KISH, L. (1965). *Survey Sampling*. New York: Wiley.
- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, Ser. A*, 139, 80-95.
- LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- LITTLE, R.J.A. (1986a). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- LITTLE, R.J.A. (1986b). Missing data in Census Bureau surveys. *Proceedings of the Second Annual Census Bureau Research Conference*, 442-454.

- LITTLE, R.J.A., and DAVID, M.H. (1983). Weighting adjustments for non-response in panel surveys. Working Paper, Washington, D.C.: U.S. Bureau of the Census.
- OH, H.L., and SCHEUREN, F. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.
- OH, H.L., and SCHEUREN, F. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 723-728.
- OH, H.L., and SCHEUREN, F. (1980). Estimating the variance impact of missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 408-415.
- OH, H.L., and SCHEUREN, F. (1983). Weighting adjustment for unit nonresponse. In *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 143-184.
- OH, H.L., SCHEUREN, F., and NISSELSOHN, H. (1980). Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 416-420.
- PALMER, S. (1967). On the character and influence of nonresponse in the Current Population Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 73-80.
- PALMER, S., and JONES, C. (1966). A look at alternate imputation procedures for CPS noninterviews. Washington, D.C.: U.S. Bureau of the Census memorandum.
- POLITZ, A., and SIMMONS, W. (1949). I. An attempt to get the 'not at homes' into the sample without callbacks. II. Further theoretical considerations regarding the plan for eliminating callbacks. *Journal of the American Statistical Association*, 44, 9-31.
- POLITZ, A., and SIMMONS, W. (1950). Note on an attempt to get the 'not at homes' into the sample without callbacks. *Journal of the American Statistical Association*, 45, 136-137.
- RUBIN, D.B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-34.
- RUBIN, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. *Bulletin of the International Statistical Institute*, 48(2), 517-532.
- RUBIN, D.B., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- SANDE, G. (1979). Numerical edit and imputation. Paper presented to the International Association for Statistical Computing, 42nd Session of the International Statistical Institute.
- SANDE, I.G. (1983). Hot-deck imputation procedures. In *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (Eds. W.G. Madow and I. Olkin), New York: Academic Press, 339-349.
- SANTOS, R.L. (1981). Effects of imputation on regression coefficients. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 140-145.
- THOMSEN, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidskrift*, 4, 278-283.
- THOMSEN, I., and SIRING, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. In *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (Eds. W.G. Madow and I. Olkin), New York: Academic Press, 25-29.
- VACEK, P.M., and ASHIKAGA, T. (1980). An examination of the nearest neighbor rule for imputing missing values. *Proceedings of the Statistical Computing Section, American Statistical Association*, 326-331.
- WELNIAK, E.J., and CODER, J.F. (1980). A measure of the bias in the March CPS earnings imputation system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 421-425.

On the Definitions of Response Rates

R. PLATEK and G.B. GRAY¹

ABSTRACT

In this paper, different types of response/nonresponse and associated measures such as rates are provided and discussed together with their implications on both estimation and administrative procedures. The missing data problems lead to inconsistent terminology related to nonresponse such as completion rates, eligibility rates, contact rates, and refusal rates, many of which can be defined in different ways. In addition, there are item nonresponse rates as well as characteristic response rates. Depending on the uses, the rates may be weighted or unweighted.

KEY WORDS: Eligibility; Completion; Contact; Refusal; Response Rates.

1. INTRODUCTION

The census or sample survey data are gathered by any one of such procedures as personal interview, telephone, or mail. It sometimes happens that some units may not respond for such reasons as "not at home", "away on vacation", "units closed", "respondent refusal", "unit vacant" or "demolished", etc. Other units may respond only partially, e.g. some but not all persons within a dwelling may respond or the units may respond to some but not all questions. Furthermore, units may respond to questions but provide incorrect or inaccurate responses.

Thus, any survey, whatever its type and method of data collection, will suffer from missing data due to nonresponse. Nonresponse has been generally recognized as an important measure of the quality of data since it affects the estimates by introducing a possible bias in the estimates and an increase in sampling variance because of the reduced sample. The relationship between sampling variance and the nonresponse rate is fairly straightforward. However, the relationship between the bias and the size of nonresponse while perhaps more important is less obvious since it depends on both the magnitude of nonresponse and the differences in the characteristics between respondents and nonrespondents. One can speculate that the nonresponse bias is proportional to the nonresponse rate. For a given response rate, the percentage bias would then be independent of sample size. However, the sampling variance is affected by the sample size and is inversely proportional to the responding sample size. Thus, the nonresponse bias may not be nearly so serious relative to the sampling errors for small samples as it is for large samples. The apparent confidence interval may cover the true value in the case of small samples but may not in the case of large samples in the presence of nonresponse bias. If we measure the "seriousness" of the nonresponse bias by the ratio of the nonresponse bias to the coefficient of sampling variation, then the "seriousness" of the nonresponse bias is proportional to the square root of the responding sample size times the nonresponse rate.

In a more practical way, the size of response/nonresponse may indicate the operational problems and provide an insight into the reliability of survey data. However, different types of response/nonresponse rates are used for these two purposes, depending upon whether or

¹ R. Platek, formerly, Director, Census and Households Survey Methods Division, Statistics Canada, G.B. Gray, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6.

or not a contact has been made with a designated unit. One can therefore distinguish between "contact" and "no contact" of types. One type such as "no one at home" or "temporarily absent" is in fact a "no contact" problem and is primarily operationally oriented. The other type is the true nonresponse problem, where contact has been made with the selected unit but no response or acceptable response is obtained.

In an interview process itself an interviewer may find units in the sample that should not be there (ineligible for the sample). Also, there will be units with questionnaires only or partially completed as well as units with all questionnaires completed. Each of these events may be defined as a rate, i.e. eligibility rate, item response rate, completion rate, etc. The distinction between the "true" nonresponse and other causes affecting the total size of nonresponse rate may give rise to different interpretations.

The interpretation of response/nonresponse rates is particularly difficult when one deals with complex survey designs since the concentration of nonresponse may be higher in one area or class than in another. Still, response rates have been used as proxies for data quality by almost all survey statisticians. That is why the interest in collecting data on nonresponse and the evaluation of it has usually been part of survey taking. However, only the measures of bias, variance, and the resultant mean square error from all sources of sampling and non-sampling errors can provide an informed basis for evaluating survey results.

Recently, nonresponse has been increasing in many surveys in Canada and elsewhere. Consequently, there is a greater need than ever before to monitor nonresponse rates, to make comparisons between surveys, countries, survey organizations, and to ensure some degree of comparability. There have been attempts to standardize the definition of response rate and its complement, the nonresponse rate; see for example, Kviz (1977), Cannell (1978). Problems of inconsistent definitions of response rates related to telephone surveys are described by Wiseman and McDonald (1980).

There are also problems of inconsistent terminology with regard to response/nonresponse in surveys. Terms such as completion rate, contact rate, and under-coverage rate have been used in different contexts in reports and articles dealing with data collection. While these terms may be readily distinguished in an individual report, they may be confusing and subject to conflicting interpretations, when studying different reports.

To consider response/nonresponse problems, a distinction must be made between unit and item nonresponse rates. Unit nonresponse rates generally pertain to the level at which survey data are gathered during the first contact. Examples of the level could be a dwelling, individual, store or establishment. However, in the case of multi-stage sampling, there may be nonresponse of all units within clusters or even primary sampling units (psu) so that unit nonresponse could apply to a selected cluster or psu as well as a dwelling or individual.

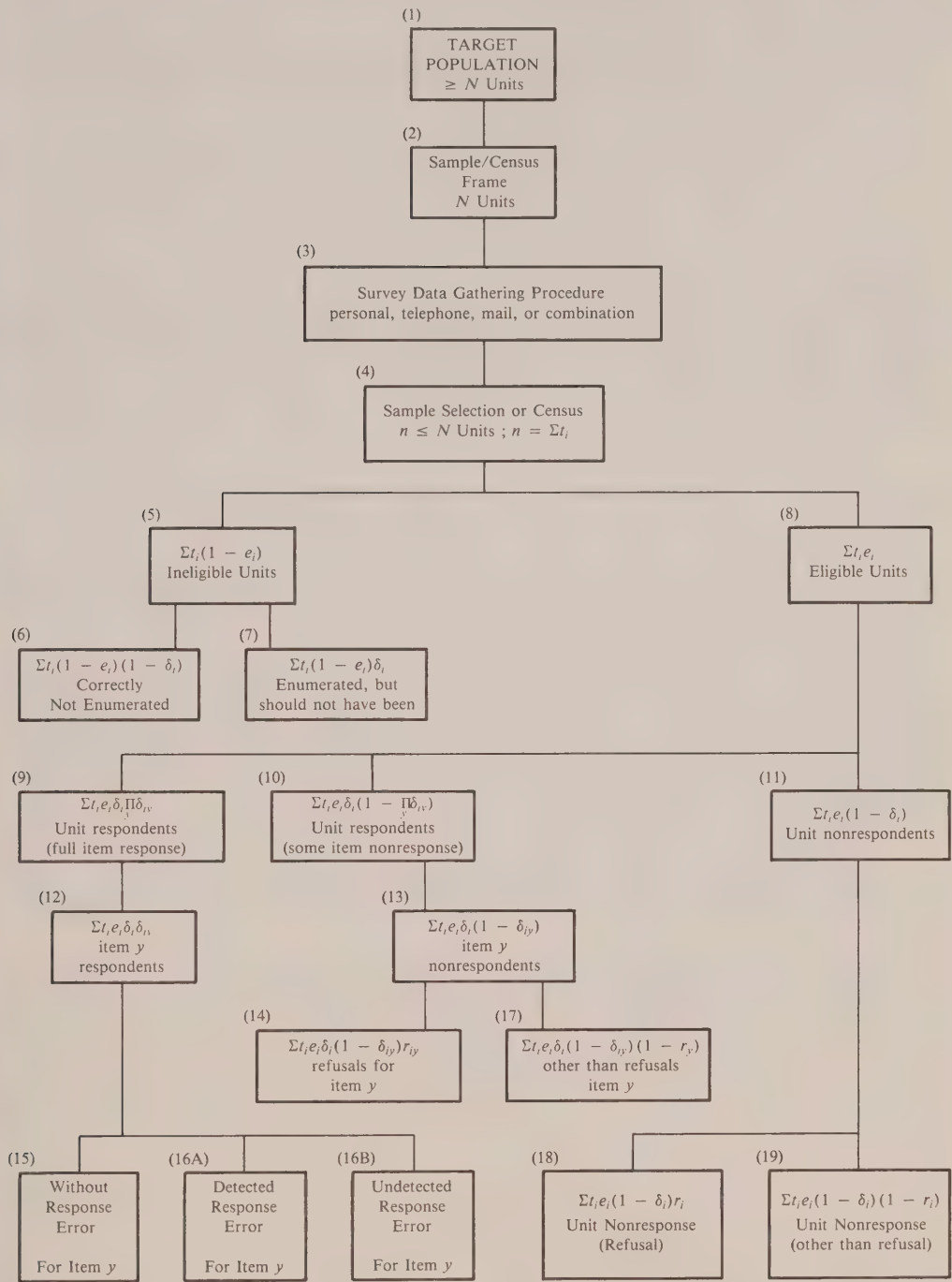
Item nonresponse usually pertains to the questionnaires, where information has been provided for some questions but not to all that should have been provided. However, if a unit fails to respond, it automatically fails to respond to any item. Hence, unit nonresponse and item nonresponse are distinct events that should be dealt with separately.

The response rates may pertain to the whole sample and part of a sample such as design-dependent areas or they may apply to administrative areas such as an interviewer assignment, or a group of assignments overseen by a supervisor or field office.

2. RESPONSE/NONRESPONSE COMPONENTS

In order to define various response rates and discuss their uses and applications, it is necessary to split up the target population for the sample or census into the various components, by type of response/nonresponse. Table 1 accomplishes this very purpose, indicating most of the important components of the whole survey that will be used in the rates. Once a target population (Box 1) is defined for a survey, a survey frame of N units (Box 2) is then determined.

Table 1
Response/Nonresponse Components



$e_i = 1,0$ (unit eligible/ineligible)
 $t_i = 1,0$ (selected/not selected)

$\delta_i = 1,0$ (unit response/nonresponse)
 $\delta_{iy} = 1,0$ (item y response/nonresponse)
 $r_i = 1,0$ according as unit refused or not
For $r_i = 0$, mainly "Not at Home"
or "Temporarily Absent"

It should be mentioned that as a result of possible under- and over-coverage of units the frame may not correspond exactly to the target population. Since under- coverage is usually more prevalent than over-coverage in practice, the actual target population usually contains more than N units.

For the survey to be taken, a data gathering procedure (Box 3) and an appropriate design are decided upon, by or census $n = \sum t_i$ units are selected, where:

$$t_i = 1 \text{ or } 0 \text{ according as unit } i \text{ is selected or not,}$$

$$\sum = \text{summation over all } N \text{ units in the survey frame.}$$

Often, in a sample frame, N may not be precisely known but rather can only be estimated from the sample. This is often the case in multi-stage probability samples with area sampling at earlier stages of selection.

Out of the sample of n units, $\sum t_i e_i$ are eligible (Box 8) and $\sum t_i (1 - e_i)$ are ineligible (Box 5) for the survey, where

$$e_i = 1 \text{ or } 0 \text{ according as unit } i \text{ is eligible or not.}$$

Sometimes the eligibility criterion may not be determined if the unit cannot be contacted while at other times the eligibility criterion is obvious from the physical appearance, such as vacant/non-vacant dwellings in a household survey.

The $\sum t_i (1 - e_i)$ ineligible units of (Box 5) may be split up between $\sum t_i (1 - e_i) (1 - \delta_i)$ units not interviewed just as they should not have been (Box 6) and $\sum t_i (1 - e_i) \delta_i$ units incorrectly interviewed (Box 7). One hopes that the number of such units in Box 7 is non-existent or at least very small. However, if such units are discovered, they should be deleted from the sample. In the above and in the breakdowns that follow, $\delta_i = 1$ or 0 according as unit i responded or did not respond.

The $\sum t_i e_i$ eligible units (Box 8) may be split up between $\sum t_i e_i \delta_i$ unit respondents (Box 9 + Box 10) and $\sum t_i e_i (1 - \delta_i)$ unit nonrespondents (Box 11), i.e. they provided no usable survey data and little, if anything, is known about the units, except perhaps their geographic location.

The $\sum t_i e_i \delta_i$ units respondents may be split up first between $\sum t_i e_i \delta_i \prod_y (\delta_{iy})$ units, free of item nonresponse, but with possible response errors (Box 9) and $\sum t_i e_i \delta_i [1 - \prod_y (\delta_{iy})]$ units with item nonresponse in at least one characteristic but not in all characteristics (Box 10). Here $\delta_{iy} = 1$ or 0 according as responding unit i responds or does not respond to item or characteristic y . In (Box 9), $\delta_{iy} = 1$ for unit i and all items while in (Box 10), $\delta_{iy} = 0$ for one or more items but not for all of them. For a particular item y , some of the $\sum t_i e_i \delta_i \delta_{iy}$ item y respondents (Box 12) come from those unit respondents, free of item nonresponse in (Box 9) while the remainder come from those unit respondents with some item nonresponse among one or more items other than item y . The $t_i e_i \delta_i (1 - \delta_{iy})$ item y nonrespondents of (Box 13) come from those unit respondents with some item nonresponse of (Box 10) that include item y .

The item y respondents of (Box 12) may be decomposed into three components, (i) those units with item y free of response error, (ii) those with a detected response error for item y , and (iii) those with an undetected response error for item y , in Boxes 15, 16A, and 16B respectively.

The $\sum t_i e_i \delta_i (1 - \delta_{iy})$ item y nonrespondents (Box 13) all come from the unit respondents, i.e. $\delta_i = 1$, $\delta_{iy} = 0$. These item nonrespondents may be decomposed into 2 components, viz., (i) those who refused to reply to question y or those who terminated the interview prior to item y (Box 14) and (ii) those who failed to reply to supply data for item y because of misunderstanding by either the respondent or interviewer or because of other reasons such as failure to follow the proper path in the questionnaire.

Finally, the unit nonrespondent (Box 11) may be split up among refusals (Box 18) and other than refusals (Box 19) mainly non-contacts with reasons such as not at home or temporarily absent. Here, $r_i = 1$ for refusal and $r_i = 0$ for cases of "other than refusal". The cases of "other than refusals" pertain mainly to "not at Home" or "Temporarily absent."

In order to count the respondents and nonrespondents according to type and reason, careful records must be kept of every sampled unit. This is essential if a probability sample is not to deteriorate into a quota sample, for example, because of ad hoc treatment of nonresponse, such as arbitrary substitution of other units for the nonrespondents. In the case of quota samples, it is sometimes difficult or impossible to distinguish substituted units from originally selected units when survey takers try to reach the quota with easy-to-obtain survey data from co-operative respondents rather than attempt call-backs of nonrespondents.

Even in probability samples with units carefully labelled and monitored according to plan, it is sometimes difficult to determine precisely the reason for nonresponse among the units that failed to be contacted. The problem is usually most straightforward in the case of personal interviews. However, even in that case, it may be difficult to distinguish "no one at home" from "temporarily absent" or "refusals" from "non-contacts" when persons are obviously at home but refuse to answer the door. In the case of telephone interviews, "no answer" or "busy signal" reveals nothing about the lack of contact of the selected unit although "refusals" of contacted units by telephone may be evident. In the case of mail surveys, when the mail is not returned, the reason could be "refusal" just as easily as "temporarily absent". The "not at home (unit)" in the usual context of nonresponse studies as distinguished from "away from home (unit)" does not apply to mail surveys. In mail surveys, the reason for nonresponse usually must be determined by personal or telephone follow-up of the unit, often by sub-sampling nonrespondents, some of which may become respondents while others may remain nonrespondents for reasons that may be determined.

The eligibility of selected units is usually evident in the case of personal interviews although failure to contact the units may result in an interviewer's inability to screen out undesirable types of units for a particular survey. No phone answers or busy signals may result in a complete failure to determine either the eligibility or type of nonresponse of the unit. Disconnected telephone numbers or ineligible telephone respondents in a screening survey will provide some measures of ineligibility in a telephone survey. In the case of mail surveys, some returned mail or addresses non-existent among selected units may yield clues about some types of ineligibility while other types may be discernable only by means of personal or telephone follow-up.

3. DEFINITIONS OF VARIOUS RATES

The sample of $n = \sum t_i$ units decomposed in Table 1 in section (2) into eligible units, unit respondents/nonrespondents, refusals, item respondents/non-respondents, etc. leads to many different types of rates which are defined below. For each rate, the numerator is a particular subset of the denominator. Wherever possible, the rate is defined in terms of the counts of units as broken down in Table 1.

(a) *Eligibility Rate*

The eligibility rate is given by:

$$\bar{e} = \sum_i t_i e_i / \sum_i t_i = (\text{Box 8})/(\text{Box 4}). \quad (3.1)$$

Wiseman and McDonald (1980) used the term "incidence rate" but applied the term only to selected persons of telephone samples that actually answered (responded) at the screening phase to determine their eligibility for the survey.

The eligibility rate, as in (3.1), demonstrates the quality of the survey design in selecting eligible units from a frame, where the eligibility may not be readily determinable without some cursory contact or observation. The rate provides, at the screening stage, information to determine how many eligible units will result at the survey data gathering stage. Thus, the rate may be employed at the design stage if data on eligibility are available from earlier studies. Depending upon the nature and procedure of the survey, the eligibility of units may not be determinable among non-contact or even refusable units. There are two alternatives to the definition of eligibility rate and response rates (which will be defined later) pertaining to eligible units. One can assume, for conservative estimates of data quality and the quality of the procedure for gathering survey data that all non-contacts and refusals would be eligible even though realistically the proportion of eligible units among such nonrespondents is often lower than among respondents and non-respondents for which the eligibility criteria are known. Under the above assumption a lower bound for the response rate and an upper bound for eligibility rate would be obtained. Alternatively, one can assume the same proportion of eligible units among units whose eligibility cannot be determined as among those whose eligibility are known. Under that assumption we would likely have a slight over-estimate of eligibility rate and some of the other rates.

(b) *Response and Completion Rates*

- (i) According to one of two alternative definitions provided by the U.S. Federal Committee on Statistical Methodology (1978), the response rate is the percentage of the eligible sample for which information (survey data) is obtained. Thus the response rate is defined as:

$$R_{(1)} = \sum_i t_i e_i \delta_i / \sum_i t_i e_i \quad (3.2)$$

$$= [(\text{Box 9}) + (\text{Box 10})]/(\text{Box 8}).$$

The above is the most commonly employed response rate in practice as it yields the percent of the sample for which some useful survey data are obtained once the ineligible units are deleted. All types of non-respondents of eligible units are included in the denominator.

The inverse of the above rate at an adjustment cell is frequently used as a weight adjustment to compensate for missing data of nonresponding units, for example, such rates are frequently use in the Canadian LFS for weight adjustments (see Platek and Gray 1985).

The above rate or its complement, the nonresponse rate, is frequently used for administrative and operational assessments of survey organizations. The rates are also used to assess interviewer's ability to contact respondents and to elect this co-operation to provide usable survey data, e.g., response/nonresponse rates by interview assignment. The non-response rate includes both refusals, which may be controlled by good public relations and diplomacy, and non-contacts, which may be beyond the control of the interviewer. Hence,

whereever possible, the nonresponse rates are frequently split up by reasons. The overall response rate in LFS is about 95% in most months. Out of the 5% nonresponse about 1% are refusals.

A similar rate to the above was defined as a completion rate by Kviz (1977), who included the whole sample in the denominator. Such a rate may provide a more conservative estimate of quality that (3.2) in that ineligible units such as vacants are included in the denominator. For example, in the LFS, the completion rate by Kviz's definition would drop from 95% according to 3.2 to about 85%.

(ii) Another definition by the above-mentioned committee is the percentage of times an interviewer obtains interviews at sample addresses, where contacts are made given by:

$$R_{(2)} = \sum_i t_i \delta_i / \sum_i t_i [\delta_i + (1 - \delta_i)r_i], \tag{3.3}$$

where unit i refused or did not refuse according as $r_i = 1$ or 0 respectively. The above was defined as a completion rate by O'Neill Groves, and Cannell (1979). If as in (3.3) the eligibility of all units that are contacted can be determined, then another and perhaps superior (known or estimated) definition of the above rate pertaining to eligible units can be given by

$$\begin{aligned} R_{(3)} &= \sum_i t_i \delta_i e_i / \sum_i t_i e_i [\delta_i + (1 - \delta_i)r_i] \\ &= [(\text{Box } 9) + (\text{Box } 10)] / [(\text{Box } 9) + (\text{Box } 10) + \text{Box } 18] \end{aligned} \tag{3.4}$$

where e_i , the eligibility criterion is defined after Table 1.

The above rates (3.3) and (3.4) may be useful in personal and telephone surveys where nonrespondents may include non-contacts and refusals. The rates are not practival in mail surveys unless there is a telephone or peronal follow-up of nonrespondents since in most pune mail surveys, the survey organization is forced with either response or nonresponse with unknown reasons. Where the above rates may be useful, however, they measure the ability of a data collection method to elect co-operation of responsible respondents at selected units, given that they are contacted. The non-contacts, that may be beyond the control of interviewers in some survey procedures are removed from the rates entirely.

The response rate in (3.4) was also defined as completion rates by Klecka and Tuchfarber (1979), who assumed, perhaps unrealistically, that all refusals were eligible for the survey. The completion rate would then have ben a conservative estimate for the measure of performance of the data collection method in eliciting the co-operation of eligible units. Alternatively, one may assume the eligibility among refusals to be the same proportion among refusals as among completed and other limits whose eligibility criteria is known.

(c) *Contact Rates*

A "contact rate", defined by Hauck (1974) is the percentage of sample units that are contacted as:

$$R_{(4)} = \frac{\text{Completed interviews} + \text{Refusals (contacted)}}{\text{Completed interviews} + \text{Refusals (contacted)} + \text{Noncontacts}}$$

where the "Noncontacts" were assumed to be eligible for a conservative estimate of the success in contacting sampled units. The "Refusals" may include "Terminations" or "Incomplete Interviews" that are essentially "Refusals" for some items as in (Box 10) of Table 1.

The algebraic expression for the contact rate is given by:

$$R_{(4)} = \frac{\sum_i t_i \delta_i e_i + \sum_i t_i (1 - \delta_i) r_i \hat{e}_i}{\sum_i t_i \delta_i e_i + \sum_i t_i (1 - \delta_i) r_i \hat{e}_i + \sum_i t_i (1 - \delta_i) (1 - r_i) \hat{e}_i} \quad (3.5)$$

$$= \frac{(\text{Box 9}) + (\text{Box 10}) + (\text{Box 18})}{(\text{Box 9}) + (\text{Box 10}) + (\text{Box 18}) + (\text{Box 19})}, \text{ where}$$

$\hat{e}_i = e_i = 1$ or 0 if eligibility criterion is known,

and, for non-contacts,

$\hat{e}_i = 1$ according to Hauck definition,

or $\hat{e}_i = \bar{e}$, the average eligibility rate among those units whose eligibility criteria are known.

The contact rate measures the ability of the survey organization or interviewers to contact respondents whether or not they succeeded in eliciting their co-operation. In the LFS, the contact rate among non-vacant dwellings is around 96% each month.

(d) *Refusal Rate (Non-refusal Rate)*

Two definitions of refusal rates are given by Hauck (1974) and Wiseman and McDonald (1980) respectively as:

$$F_1 = \frac{\text{number of refusals}}{\text{number of completed interviews and refusals}}$$

$$= \frac{\sum_i t_i \hat{e}_i (1 - \delta_i) r_i}{\sum_i t_i e_i \delta_i + \sum_i t_i \hat{e}_i (1 - \delta_i) r_i} \quad (3.6)$$

$$= (\text{Box 18}) / [(\text{Box 9}) + (\text{Box 10}) + (\text{Box 18})] = 1 - R_{(3)}.$$

and

$$F_2 = \frac{\text{number of refusals}}{\text{number of all selected units}}$$

$$= \frac{\sum_i t_i (1 - \delta_i) r_i}{\sum_i t_i} \quad (3.7)$$

$$= (\text{Box 18}) / (\text{Box 4}).$$

With the eligibility criteria taken into account, the refusal rate in (3.7) may be given by:

$$F_3 = \frac{\sum_i t_i \hat{e}_i (1 - \delta_i) r_i}{\sum_i t_i \hat{e}_i} \quad (3.8)$$

$$= (\text{Box 18}) / (\text{Box 8}), \text{ where } \hat{e}_i \text{ is defined after (3.5).}$$

The refusal rate measures the extent of the inability of the survey organization or the interviewer to elicit the co-operation of units to provide usable survey data, relative to all contacted units (3.6), relative to the whole sample (3.7) or relative to the eligible sample (3.8). In (3.6), one may wish to determine a "pure" refusal rate without non-contacts that are often beyond the interviewers' control in order to study the efficiency of a questionnaire or effect of the survey topic on the co-operation of contacted units. Alternatively, in (3.7) and (3.8), one may prefer to examine the refusals rate as one, of several components of overall nonresponse.

(e) *Item Response/Nonresponse Rates*

Complex questionnaire design may result in item nonresponse of specific questions for reasons other than refusals, as noted in Box 17. A controversial or personal question or termination of the interview may result in a refusal to provide data for a specific item as in (Box 14).

Thus, one may measure the overall item nonresponse rate for item y , relative to all responding units, given by:

$$R_y = \frac{(\text{Box 13})}{(\text{Box 9}) + (\text{Box 10})}$$

or if item y is relevant only for some units (questionnaires) but not for all of them, one may measure the item nonresponse relative to only those responding units for which item y is relevant (eligible). Consequently, one may define a whole set of item response/nonresponse/eligibility rates, analogous to the unit rates replacing in the rates the number of units (eligible/ineligible)/(responding/refusing, etc.) with the number of responding units (eligible or relevant for item y , irrelevant, responding for item y /refusing for item y etc.) respectively. Most of the rates pertaining to units other than contact rates should have their item y counterparts readily defined by making the proper substitutions in the expressions. However, it may be more difficult to record the reasons for item nonresponse, compared with unit nonresponse, as frequently the item nonresponse is detected only through an edit and imputation routine.

(f) *Weighted Rates and Characteristic Rates*

In the case of sample with different sample weights Π_i^{-1} 's for the units as in probability proportional to size (*pps*) sampling, all of the above rates may be defined as weighted rates by applying the sample weight Π_i^{-1} with the sample selection indicator variable t_i in all the expressions. In the case of self-weighting samples in an area or class for which the rates are calculated the sample weights are redundant. In *pps* sampling at the final stage, however, the usual tendency is for large units to respond more readily than small ones so that weighted response rates, with smaller sample weights applied to the large units than for small units, tend to be smaller than unweighted rates based on the counts of units as in Table 1.

The weighted response rates estimate the proportion of the population that would have responded to the survey under similar survey conditions while the unweighted response rates provide a measure of data collection performance only for the sample or sub-sample pertaining to a specified area or class.

By estimating the nonresponse rate for the entire population rather than for the sample as the unweighted rates do, the weighted rate may provide misleading information on the quality of the data since it may distort the distribution of characteristics in the sample. The advantage of the weighted rates, however, is that the units are added to population levels

rather than sample levels so that one obtains an estimate of the rate that would prevail at census levels under similar conditions of gathering survey data. The weighted response rates may under some circumstances be used as weight adjustment factors to inflate the respondents to the full sample in adjustment cells.

When defining characteristic response rates factors include the observed response y_i among item respondents, the imputed value z_{iy} for item nonresponse and the imputed value for z_i for unit nonresponse, which is usually the mean of the respondents in an adjustment cell. If some auxiliary value X_i is known for all units, whether or not they respond, then a characteristic x response rate may be readily calculated and used as a weight adjustment when x is highly correlated with y . The characteristic y response rate, weighted by Π_i^{-1} or unweighted, may be useful in studying the potential nonresponse bias by comparing the characteristic y response rates with the weighted or unweighted response rates based on counts of units.

4. FINAL REMARKS

Standardization of the definitions of the rates appears to be difficult, owing to the variety of uses and studies of nonresponse and owing to the careful record keeping demanded of survey takers. As long as the rates are unambiguously defined and appropriately applied in their analysis standard definitions for all types of surveys and survey data gathering procedures, may not be all that important. However, in each particular case, the rate should be carefully defined with clear demonstration of the purpose for which it is intended and the reason why it is adopted.

Another issue of standardization dealing with the topic of response/non-response rates is the standard of what is expected from past experience for given surveys, type of survey, subject matter and interview procedure. For example, the response rate, according to 3.2, in the LFS, is expected to be in the 93 to 95% range, with slightly lower rates in the summer months. Out of the 5 to 7% nonresponse, 1% or so may be expected to be refusals. The overall rates have been remarkably consistent for the history of the survey.

It has been observed (see Platek 1977) that finance-oriented surveys tend to have lower response (higher nonresponse) rates than surveys dealing with other topics. The finance surveys appear to be around 25% nonresponse while most of the others centre around 10 to 15%. Also, telephone surveys appear to have a slightly higher nonresponse rate (by about 2 to 3%) than personal surveys for similar subject matter. Thus, from experience, one can determine a standard objective for surveys of a given subject and interview procedure.

It has been observed in publications such as Wiseman and McDonald (1980) that there are many opinions of the way nonresponse should be defined and measured. Thus, it appears that one must grapple with the alternative definitions and terms and obtain relationships between them under various survey conditions. We have attempted to focus on the problems of the various definitions, terms and standards of response rates but have not solved the problems. A proper study can really be undertaken only with a thorough evaluation of survey records, which is possible only when good records are kept. Often, particularly in the case of quota samples, in telephone and mail surveys, nonrespondents are set aside and other units are substituted for them and treated like the originally selected units. The result is a higher observed quality of survey than is the case in reality because of the hidden nonresponse bias. Consequently, the way of treating nonrespondents and the evaluation of nonresponse, completion, etc. must be planned in advance of the survey data gathering in order to deal with it properly rather than during or after the survey.

REFERENCES

- CANNELL, CHARLES (1978). Discussion of response rates. Health Survey Research Methods Conference, DHEW Publication No. (PHS) 79-3207.
- HAUCK, MATTHEW (1974). Planning field operations. In *Handbook of Marketing Research* (Robert Ferber), New York: McGraw-Hill, 147-159.
- KALTON, GRAHAM (1981). Compensating for missing survey data. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.
- KLECKA, W.R. and A.J. TUCHFARBER (1979). Random digit dialing: A comparison to personal surveys. *Public Opinion Quarterly* (Spring), 105-114.
- KVIZ, FREDERICK J. (1977). Toward a standard definition of response rate. *Public Opinion Quarterly* (Summer), 265-267.
- LINDSTRÖM, HAKAN (1983). Non-response errors in sample surveys. Urval, Nummer 16, Skriftserie utgiven av Statistiska Centralbyran, Statistics Sweden, Stockholm.
- O'NEILL, MICHAEL J., GROVES, ROBERT M., and CANNELL, CHARLES F. (1979). Telephone interview introductions and refusal rates: Experiments in increasing respondent cooperation. Paper presented at the 1979 Meeting of the American Statistical Association, Washington, D.C.
- PLATEK, RICHARD (1977). Some factors affecting non-response. Paper presented at the International Statistical Institute, New Delhi, December.
- PLATEK, RICHARD and GRAY, G.B. (1985). Some aspects of nonresponse adjustment. *Survey Methodology*, 11, 1-14.
- WISEMAN, FREDERICK, and PHILIP McDONALD (1978). The nonresponse problem in consumer telephone survey. Report No. 78-116, Marketing Science Institute, Cambridge, Mass.
- WISEMAN, FREDERICK, and PHILIP McDONALD (1980). Toward the development of industry standards for response and nonresponse rates. Report no. 80-101, Marketing Science Institute, Cambridge, Mass.

Some Optimality Results in the Presence of Nonresponse

V.P. GODAMBE and M.E. THOMPSON¹

ABSTRACT

Using the optimal estimating functions for survey sampling estimation (Godambe and Thompson 1986), we obtain some optimality results for nonresponse situations in survey sampling.

KEYWORDS: Optimum estimating function; Nonresponse.

1. INTRODUCTION AND BACKGROUND

A typical survey sampling set-up consists of a survey population \mathbf{P} of N labelled individuals i ; $\mathbf{P} = \{i: i = 1, \dots, N\}$. With each individual i is associated a real value y_i . The vector $\mathbf{y} = (y_1, \dots, y_N)$ is called the population vector. Any subset s of \mathbf{P} is called a sample. Let $S = \{s\}$. Any probability distribution p on S is called a sampling design. A sample s is drawn using a sampling design p , and the values $y_i: i \in s$ are ascertained through a survey.

Thus the data here are x_s where

$$x_s = \{s, (i, y_i): i \in s\}. \quad (1.1)$$

On the basis of the data x_s one tries to estimate a survey population parameter θ_N , that is a specified real function of the population vector \mathbf{y} ; $\theta_N = \theta_N(\mathbf{y})$.

In relation to the above estimation problem we assume a superpopulation model under which y_1, \dots, y_N are independent and for certain known covariate values $x_i, i = 1, \dots, N$,

$$\epsilon(y_i - \theta x_i) = 0, i = 1, \dots, N, \quad (1.2)$$

ϵ being the expectation with respect to the model. In the model (1.2), θ is the usual unknown regression parameter, the expectation being taken holding x_i fixed. The usual intercept term of the regression model is not mentioned in (1.2), for this term can often be eliminated by an appropriate stratification (Godambe 1982). Note the model (1.2) does not specify the variance function.

Following Godambe and Thompson (1986), for some *specified* numbers $\alpha_i, i = 1, \dots, N$, we define the survey population parameter θ_N as the solution of the equation

$$\tilde{g} = \sum_{i=1}^N (y_i - \theta x_i) \alpha_i = 0. \quad (1.3)$$

¹ V.P. Godambe and M.E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada, N2L 3G1.

That is,

$$\theta_N = \sum_{i=1}^N y_i \alpha_i / \sum_{i=1}^N x_i \alpha_i. \quad (1.4)$$

The parameter θ_N is related to the model (1.2) through the equation

$$\epsilon \bar{g} = 0. \quad (1.5)$$

Any real function h of the data χ_s in (1.1) and the parameter θ is called an *unbiased estimating function* for both the parameters θ_N and θ if

$$E(h - \bar{g}) = 0 \text{ for all } \mathbf{y} \text{ and } \theta \quad (1.6)$$

' E ' being the expectation under the sampling design p employed to draw the sample s . Because of (1.5) and (1.6) we say the solution of the equation

$$h(\chi_s, \theta) = 0,$$

for the given data χ_s , estimates both the parameters and θ and θ_N , given by (1.2) and (1.4) respectively. For the function \bar{g} in (1.4), under the sampling design p , let $H_{(p)}$ be the class of all unbiased estimating functions h . That is

$$H(p) = \{h: E(h - \bar{g}) = 0 \text{ for all } \mathbf{y} \text{ and } \theta\}. \quad (1.7)$$

Now we say an *estimating function* $h^* \in H(p)$ is *optimum* if

$$\epsilon E(h^*)^2 \leq \epsilon E(h)^2, \text{ for all } h \in H(p) \quad (1.8)$$

(Godambe and Thompson 1986). Further, when the inequality (1.8) is satisfied,

$$h^* = 0 \quad (1.9)$$

is said to be the *optimum estimating equation* for estimating the parameter θ_N given by (1.3) and (1.4).

For the sampling design p , used to draw a sample s , let π_i , $i = 1, \dots, N$ be the inclusion probabilities. That is

$$\pi_i = \sum_{s \ni i} p(s), \quad i = 1, \dots, N, \quad (1.10)$$

where $s \ni i$ indicates all samples s which include the individual i . We assume

$$\pi_i > 0, \quad i = 1, \dots, N. \quad (1.11)$$

Theorem 1.1. (Godambe and Thompson 1986). For any sampling design p satisfying (1.11), under the model (1.2), in the class of all unbiased estimating functions $H(p)$ in (1.7), the optimum h^* , that is h^* satisfying (1.8), is given by

$$h^* = \sum_{i \in s} (y_i - \theta x_i) \alpha_i / \pi_i, \quad (1.12)$$

π_i being the inclusion probability given by (1.10). Thus the optimum estimating equation here is

$$\sum_{i \in s} (y_i - \theta x_i) \alpha_i / \pi_i = 0. \quad (1.13)$$

The estimate $\hat{\theta}_s$ of the survey population parameter θ_N in (1.4) and the superpopulation parameter θ in (1.2) is given by

$$\hat{\theta}_s = \frac{\sum_{i \in s} y_i \alpha_i / \pi_i}{\sum_{i \in s} x_i \alpha_i / \pi_i}. \quad (1.14)$$

This estimate was previously put forward by Brewer (1963) and Hájek (1971) on some "plausibility" considerations.

To explain the relationships of Theorem 1.1 above with earlier optimality results (e.g. Godambe 1982) we put $\alpha_i \equiv 1$ in (1.3) and therefore in (1.2). Further, we consider a superpopulation model obtained from (1.2) by letting $\theta = \theta_0$, a specified value. Now for any sampling design with inclusion probabilities π_i satisfying (1.11), in the class of all design unbiased estimates of θ_N (in (1.4) with $\alpha_i = 1$, $i = 1, \dots, N$), the superpopulation expectation of the design variance is minimized for the estimate

$$e = \frac{1}{X} \left\{ \sum_{i \in s} \frac{y_i - \theta_0 x_i}{\pi_i} + \theta_0 \sum_{i=1}^N x_i \right\} \quad (1.15)$$

where $X = \sum_{i=1}^N x_i$. This "optimality" of the estimate e at $\theta = \theta_0$ carries over to all values of θ if the sampling design is such that

$$\text{Probability} \left\{ s: \left(\sum_{i \in s} \frac{x_i}{\pi_i} - \sum_{i=1}^N x_i \right) = 0 \right\} = 1. \quad (1.16)$$

Now when the sampling design satisfies condition (1.16), then $\hat{\theta}_s$ in (1.14) is equal to e in (1.15). Thus all the earlier optimality results are covered by Theorem 1.1, and it does a great deal more: in many situations, such as for designs with $\pi_i \propto x_i$, the condition (1.16) implies a *fixed sample size* design. In contrast the "optimality" in Theorem 1.1 holds regardless of the fixed sample size design condition. That is, the "optimality" is available for *random sample size* designs, which are common in the nonresponse situations discussed subsequently.

2. NONRESPONSE AND OPTIMALITY

Suppose a sample s is drawn from the survey population \mathbf{P} , using a sampling design p . Suppose because of nonresponse the variate values y_i are available only for the subset $s' \subset s$; $s - s'$ are the non-respondents. Thus now the data instead of χ_s in (1.1) are

$$\chi_{s,s'} = (s, s', \{ (i, y_i) : i \in s' \}). \quad (2.1)$$

We may now consider two problems of estimation:

- (I) If there were no nonresponse, that is if all the data χ_s in (1.1) were available, we would have estimated the survey population parameter θ_N in (1.4) by solving the optimum estimating equation given by (1.12), namely $h^* = 0$. When the hypothetical data χ_s are replaced by $\chi_{s,s'}$ in (2.1), one may try to estimate h^* with some function $h'(\chi_{s,s'})$. This is in line with a suggestion of Rubin (1976). Following (1.7) we define the class of unbiased estimating functions h' (for h^* , given the sample s) as

$$H'(p,.,s) = \{h' : E(h' - h^*|s) = 0, \text{ for all } \mathbf{y} \text{ \& } \theta\}; \quad (2.2)$$

the '.' in H' indicates that the class H' would be specified only after the *response mechanism* is specified. Again we define h'^* as the optimum estimating function in H' in (2.2), if $h'^* \in H'$ and if under the model (1.2), $\epsilon E(h'^*)^2 \leq \epsilon E(h')^2$ for all $h' \in H'$.

- (II) Alternatively we could try to estimate the survey population parameter θ_N directly, that is without estimating h^* as in (I) above, from the data $\chi_{s,s'}$. In line with (1.7) we define the class of unbiased estimating functions $h''(\chi_{s,s'})$:

$$H''(p,.) = \{h'' : E(h'' - \bar{g}) = 0, \text{ for all } \mathbf{y} \text{ \& } \theta\}; \quad (2.3)$$

as before the '.' in H'' indicates that the class H'' , for its specification, requires the specification of the *response mechanism*. Again h''^* is called the *optimum estimating function* in H'' if $h''^* \in H''$ and if under (1.2), $\epsilon E(h''^*)^2 \leq \epsilon E(h'')^2$ for all estimating functions $h'' \in H''$.

In $H'(p,.,s)$ and $H''(p,.)$ of (2.2) and (2.3) we have left the response mechanism '.' unspecified. Now we specify it.

RESPONSE MECHANISM: If the individual ' i ' of the survey population \mathbf{P} were included in the sample s drawn,

$$\begin{aligned} &\text{'i' would respond with known probability } q_i \\ &\text{and would fail to respond with probability } 1 - q_i, \end{aligned} \quad (2.4)$$

$i = 1, \dots, N$; we assume $q_i > 0$, $i = 1, \dots, N$.

The response mechanism $\mathbf{q} = (q_1, \dots, q_N)$ in (2.4) completely characterizes the class $H'(p,.,s)$ in (2.2) as $H'(p, \mathbf{q}, s)$ and $H''(p,.)$ in (2.3) as $H''(p, \mathbf{q})$.

The case (I) above is implemented by the following Theorem 2.1 and the remaining Theorems 2.2, 2.3 and 2.4 implement the case (II).

Theorem 2.1. For any sampling design p satisfying (1.11), and for any sample s , in the class of estimating functions $H'(p, \mathbf{q}, s)$ in (2.2) under the superpopulation model (1.2) $\epsilon E\{h'\}^2 | s\}$ is minimized for $h' = h'^*$ where

$$h'^* = \sum_{i \in s'} (y_i - \theta x_i) \alpha_i / \pi_i q_i; \tag{2.5}$$

that is h'^* is the optimum estimating function in $H'(p, \mathbf{q}, s)$. \square

Proof. As was emphasized in Section 1, the optimality of h^* in (1.12) obtains even for random sample size designs and for any values of α_i , $i = 1, \dots, N$ in (1.3). Thus the proof of Theorem 2.1 is accomplished by replacing, in Theorem 1.1, the population 'P' by 's' and α_i by α_i / π_i , $i \in s$ and noting that now the inclusion probabilities are q_i , $i \in s$. \square

Theorem 2.2. Let \bar{H}'' be the subclass of H'' in (2.3) such that any estimating function $h''(\chi_{s,s'})$ in \bar{H}'' depends on (s, s') only through s' . Then for any sampling design p satisfying (1.11), in the class $\bar{H}''(p, \mathbf{q})$, under the superpopulation model (1.2), $\epsilon E\{(h'')^2\}$ is minimized for $h'' = h''^*$ where

$$h''^* = \sum_{i \in s'} (y_i - \theta x_i) \alpha_i / \pi_i; \tag{2.6}$$

that is h''^* is the optimum estimating function in $\bar{H}''(p, \mathbf{q})$. \square

Proof. This follows directly from Theorem 1.1, by replacing in it s by s' and the inclusion probabilities by π_i by $\pi_i q_i$, $i = 1, \dots, N$.

Theorem 2.3. The estimating function h''^* in (2.6) is the optimum estimating function in the entire class $H''(p, \mathbf{q})$ given by (2.3). That is the result of the Theorem 2.2 is valid without the restriction to the subclass \bar{H}'' of H'' . \square

Proof. For any given response probabilities \mathbf{q} in (2.4) and the sampling design p , the statistic $\{(y_i, x_i) : i \in s'\}$ is sufficient for the population vector \mathbf{y} . More specifically, referring to (1.1) and (2.1), we have the conditional probability $\text{Prob}(\chi_{s,s'} | \chi_{s'}, \mathbf{y})$ independent of \mathbf{y} . Hence for any estimating function $h'' \in H''(p, \mathbf{q})$ in (2.3) we have the estimating function $E(h'' | \chi_{s'}) = \bar{h}'' \in \bar{H}''$ and $\epsilon E(\bar{h}'')^2 \leq \epsilon E(h'')^2$. This proves Theorem 2.3.

When $s \equiv s'$, that is when there are no nonrespondents, do we still estimate h^* by $h'^* = h''^*$? The obvious negative answer to this question is obtained, as shown by Godambe (1986), by an appropriate conditioning. The same reservation tends to be felt for cases where there are only a few nonrespondents, and again appropriate conditioning holds some promise of a resolution. In summary the formal optimality of $h'^* = h''^*$ suggests that it is useful, and is likely to give good estimation when nonresponse is considerable and the relative values of the q_i are known. However, it can clearly be improved upon in situations when nonresponse is rare; improved versions will have natural conditional interpretations. Appropriate conditioning becomes even more important in the case of unknown response probabilities, as will be seen next.

Now we assume that the survey population \mathbf{P} is divided into k strata \mathbf{P}_j , of sizes N_j , $j = 1, \dots, k$. Further suppose that the response probabilities are constant within each stratum. That is

$$q_i = q^{(j)} \text{ for all } i \in \mathbf{P}_j; j = 1, \dots, k. \tag{2.7}$$

Unlike in (2.4), where the response probabilities were assumed to be known, now we assume that in (2.7), the response probabilities $q^{(j)}$, $j = 1, \dots, k$ are *unknown*. Let p_0 denote the stratified sampling design, consisting of drawing from the stratum \mathbf{P}_j , a simple random sample (without replacement) of size n_j , $j = 1, \dots, k$. Now as in (2.3) we define the class of unbiased estimating functions $h_1(\chi_{s,s'})$

$$H_i(p_0) = \{h_i: E(h_1 - \bar{g}) = 0 \text{ for all } \mathbf{y}, \theta \text{ and } q^{(j)}, j = 1, \dots, k\}, \quad (2.8)$$

where $q^{(j)}$ are as in (2.7). Let $s'_j = s' \cap \mathbf{P}_j$ and $|s'_j| = n'_j$, that is the size of the sample of respondents from the stratum \mathbf{P}_j , $j = 1, \dots, k$.

Theorem 2.4. For the sampling design p_0 , in the class of estimating functions $H_i(p_0)$ in (2.8), under the superpopulation model (1.2), $\epsilon E(h_1^2)$ is minimized for $h_1 = h_1^*$ where

$$h_1^* = \sum_{j=1}^k \sum_{i \in s'_j} (y_i - \theta x_i) \alpha_i / \left(\frac{n'_j}{N_j}\right); \quad (2.9)$$

that is h_1^* is the optimum estimating function in $H_i(p_0)$.

Proof. The sampling distribution of the data $\chi_{s,s'}$ in (2.1) depends, in addition to the unknown population vector \mathbf{y} , on the unknown (parameter) $q^{(j)}$, $j = 1, \dots, k$. Now for every fixed \mathbf{y} , the statistic n'_j , $j = 1, \dots, k$ is *completely sufficient* for the parameter $q^{(j)}$, $j = 1, \dots, k$. Hence for a fixed \mathbf{y} and θ , in (2.8),

$$\begin{aligned} [E(h_1 - \bar{g}) = 0, \text{ for all } q^{(j)}, j = 1, \dots, k] \\ \Rightarrow E\{(h_1 - \bar{g}) | n'_j, j = 1, \dots, k\} = 0, \end{aligned} \quad (2.10)$$

ignoring sets of '0' measure. Further, *conditional* on the number of respondents n'_j from the stratum \mathbf{P}_j , the probability of $i \in s'_j$ is $(n_j/N_j)(n'_j/n_j) = (n'_j/N_j)$. Hence for any estimating function $h_1 \in H_1$ in (2.8) we have from Theorem 2.3.

$$\epsilon E((h_1^*)^2 | n'_j, j = 1, \dots, k) \leq \epsilon E\{(h_1)^2 | n'_j, j = 1, \dots, k\}, \quad (2.11)$$

h_1^* being given by (2.9). Theorem 2.4 is proved by taking the expectations of both sides of (2.11) for the variations of n'_j , $j = 1, \dots, k$.

The optimum estimating function h_1^* in (2.9) has the following intuitive interpretation. If in (2.7), the response probabilities $q^{(j)}$, $j = 1, \dots, k$ were *known*, by Theorem 2.3, the optimum estimating function, for the sampling design p_0 , would be given by

$$h'' = \sum_{j=1}^k \sum_{i \in s'_j} (y_i - \theta x_i) \alpha_i / \left(\frac{n_j}{N_j} q^{(j)}\right).$$

Now when $q^{(j)}$ are unknown (which is the case in Theorem 2.4), we *estimate* them by (n'_j/n_j) , $j = 1, \dots, k$. Substituting these estimates for $q^{(j)}$ in h'' yields the estimating function h_1^* of (2.9).

These estimates obtained by solving the equations $h'^* = 0$, $h''^* = 0$ and $h_1^* = 0$ in (2.5), (2.6) and (2.9) respectively have previously been proposed, on plausibility considerations, by several authors. A good reference in this connection in Cassel et al. (1983). The assumption (2.4) of "response probabilities" seems to have evolved gradually in the literature. An interesting early reference in this connection is Hartley (1946).

3. OPTIMAL INCLUSION PROBABILITIES

It should be emphasized here that the “optimality” of the estimating function h^{**} in (2.6) was established under the superpopulation model (1.2), which does *not* specify the variance function. However the specification of the variance function in the model (1.2) would be required to obtain the “optimal” inclusion probabilities. We assume

$$\epsilon(y_i - \theta x_i)^2 = \sigma^2 f(x_i), \quad i = 1, \dots, N, \tag{3.1}$$

where f is a *known* function of x , and σ^2 can be unknown. Now for the estimating function h^{**} in (2.6), (3.1), we have

$$\epsilon E(h^{**})^2 = \sum_{i=1}^N \frac{\epsilon(y_i - \theta x_i)^2 \alpha_i^2}{\pi_i q_i} = \sigma^2 \sum_{i=1}^N \frac{f(x_i) \alpha_i^2}{\pi_i q_i} \tag{3.2}$$

In (3.2), the response probabilities q_i as said in (2.4) are given (fixed) numbers. However, (a sampling design with) the optimal inclusion probabilities can be obtained by minimizing $\epsilon E(h^{**})^2$ in (3.2) under a restriction, either (A) or (B).

$$\begin{aligned} (A): \quad & \sum_{i=1}^N \pi_i = \text{constant}, \\ (B): \quad & \sum_{i=1}^N \pi_i q_i = \text{constant} \end{aligned} \tag{3.3}$$

In (A) we hold the average size of the sample s fixed, for $E|s| = \sum_i^N \pi_i$. In (B) we hold fixed the average size of the effective sample s' , for $E|s| = \sum_i^N \pi_i q_i$. Now since the q_i are fixed numbers we have for minimizing $\epsilon E(h^{**})^2$ in (3.2), respectively,

$$\begin{aligned} (A): \quad & \pi_i \propto \left\{ \frac{f(x_i)}{q_i} \right\}^{1/2} \alpha_i, \\ (B): \quad & \pi_i \propto \frac{(f(x_i))^{1/2}}{q_i} \alpha_i. \end{aligned} \tag{3.4}$$

Denoting by n' the size of the effective sample s' , that is $|s'| = n'$, we have from (B) in (3.4),

$$\pi_i = \frac{(f(x_i))^{1/2} \alpha_i}{\{\sum_1^N (f(x_i))^{1/2} \alpha_i\}} \cdot \frac{E(n')}{q_i}, \quad i = 1, \dots, N. \tag{3.5}$$

Further for a fixed sample size design such that

$$\text{Probability } \{s: |s| \neq n\} = 0,$$

we have from (3.5).

$$\sum_{i=1}^n \pi_i = n = \sum_{i=1}^N \frac{(f(x_i))^{1/2} \alpha_i}{\{\sum_1^N (f(x_i))^{1/2} \alpha_i\}} \cdot \frac{1}{q_i} E(n'). \tag{3.6}$$

As a special case, when all the response probabilities q_i , $i = 1, \dots, N$ are equal, $q_i = q$ say, $i = 1, \dots, N$, in (3.6),

$$n = E(n')/q; \quad (3.7)$$

for instance if $q = 1/2$, the sample size of the (initial) sample s should be double the expectation of the effective sample (s') size!

Now we assume the survey population \mathbf{P} to be divided into strata P_j , $j = 1, \dots, k$ so that the response probabilities in each stratum are constant, that is they satisfy (2.7). For a stratified sampling design consisting of drawing a sample of size n_j from the stratum \mathbf{P}_j , $j = 1, \dots, k$ we have from (3.5).

$$n_j = \frac{E(n')}{q^{(j)}} \frac{\sum_{i \in \mathbf{P}_j} (f(x_i))^{1/2} \alpha_i}{\sum_{i \in \mathbf{P}} (f(x_i))^{1/2} \alpha_i}, \quad j = 1, \dots, k.$$

If $(f(x_i))^{1/2} \alpha_i$ are constant for $i = 1, \dots, N$, it is clear from (3.8) that optimal allocation implies drawing a relatively larger sample from the stratum with smaller response probability. Actually in this situation

$$E(n'_j) = E(n')/k$$

where n'_j is the size of the effective sample s'_j from the stratum \mathbf{P}_j , $j = 1, \dots, k$.

REFERENCES

- BREWER, K.R.W. (1963). Ratio estimation in finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.
- CASSEL, C.M., SARNDAL, C.E., and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problems. In *Incomplete Data in Sample Surveys*, Vol. 3, (Eds. W.G. Madow and Ingram Olkin), New York: Academic Press, 143-160.
- GODAMBE, V.P. (1982). Estimation in survey sampling: Robustness and optimality. *Journal of the American Statistical Association*, 77, 393-403.
- GODAMBE, V.P. (1986). Quasi-score function, quasi-observed Fisher information and conditioning in survey sampling (unpublished).
- GODAMBE, V.P. and THOMPSON, M.E. (1986). parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Institute Review* (to Appear).
- HAJEK, J. (1971). Contribution to discussion of paper by D. Basu. In *Foundations of Statistical Inference*, (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 236.
- HARTLEY, H.O. (1946). Discussion of paper by F. Yates. *Journal of the Royal Statistical Society Series A*, 109, 37.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-589.

Basic Ideas of Multiple Imputation for Nonresponse

DONALD B. RUBIN¹

ABSTRACT

Multiple imputation is a technique for handling survey nonresponse that replaces each missing value created by nonresponse by a vector of possible values that reflect uncertainty about which values to impute. A simple example and brief overview of the underlying theory are used to introduce the general procedure.

KEY WORDS: Survey nonresponse; Proper imputation methods; Multiple imputation.

1. INTRODUCTION

Any statistician with experience in the field of surveys knows that essentially every survey suffers from some nonresponse. That is, in practical surveys, some items in the survey instrument are not answered by all units included in the survey. Commonly, the items likely to be unanswered are the more sensitive ones, such as those concerning personal income. Because nonresponse creates missing values, the complete-data statistics that would have been used in the absence of nonresponse can no longer be calculated. An obvious desire of both the data collector and the data analyst is to get rid of the missing values and thereby restore the ability to use standard complete-data methods to draw inferences.

1.1 Imputation

It is not surprising, therefore, that a very common method of handling the missing values created by nonresponse is to fill them in, or impute them. That is, when using imputation to handle nonresponse each missing value is replaced with a real value. Many different procedures have been proposed for imputation, for instance, filling in the respondents' mean for that variable or a value predicted from the modelling of the missing variable given observed variables using respondent data; as a specific example, when the missing value is personal income, a linear regression model predicting $\log(\text{income})$ from demographic characteristics such as age, sex, education and occupation might be regarded as reasonable.

1.2 Advantages and Disadvantages of Single Imputation

In addition to the obvious advantage of allowing complete-data methods of analysis, imputation by the data collector (e.g. the Census Bureau) also has the important advantage of being able to utilize information available to the data collector but not available to an external data analyst such as a university social scientist analyzing a public-use file. This information may involve detailed knowledge of interviewing procedures and reasons for nonresponse that are too cumbersome to place in public-use files, or may be facts, such as street addresses of dwelling units, that cannot be placed on public-use files because of confidentiality constraints. This kind of information, even though inaccessible to the user of a public-use file, can often narrow the possible range of imputed values.

¹ Donald B. Rubin, Department of Statistics, Harvard University, Science Center, One Oxford Street, Cambridge, Massachusetts, 02138, U.S.A.

Just as there are obvious advantages to imputing one value for each missing value, there are obvious disadvantages of this procedure arising from the fact that the one imputed value cannot itself represent any uncertainty about which value to impute: If one value were really adequate, then that value was never missing. Hence, analyses that treat imputed values just like observed values generally systematically underestimate uncertainty, even assuming the precise reasons for nonresponse are known. Equally serious, single imputation cannot represent any additional uncertainty that arises when the reasons for nonresponse are not known.

1.3 Multiple Imputation to the Rescue

Multiple imputation, first proposed in Rubin (1977, 1978), retains the two major advantages of single imputation and rectifies its major disadvantages. As its name suggests, multiple imputation replaces each missing value by a vector composed of $M \geq 2$ possible values. The M values are ordered in the sense that the first components of the vectors for the missing values are used to create one completed data set, the second components of the vectors are used to create the second completed data set and so on. The first major advantage of single imputation is retained with multiple imputation, since standard complete-data methods are used to analyze each completed data set. The second major advantage of imputation, that is, the ability to utilize data collectors' knowledge in handling the missing values, is not only retained but actually enhanced. In addition to allowing data collectors to use their knowledge to make point estimates for imputed values, multiple imputations allow data collectors to reflect their uncertainty as to which values to impute. This uncertainty is of two types: sampling variability assuming the reasons for nonresponse are known, and variability due to uncertainty about the reasons for nonresponse. Under each posited model for nonresponse, two or more imputations are created to reflect sampling variability under that model; imputations under more than one model for nonresponse reflect uncertainty about the reasons for nonresponse. The multiple imputations within one model are called repetitions and can be combined to form a valid inference under that model; the inferences under different models can be contrasted to reveal sensitivity of answers to posited reasons for nonresponse.

Before reviewing some more general results in Section 3, Section 2 illustrates essential ideas in a highly artificial example used in Rubin (1986a), which is a comprehensive treatment of multiple imputation. Other references on multiple imputation include Rubin (1979, 1980, 1986b), Herzog and Rubin (1983), Li (1985), Schenker (1985), Rubin and Schenker (1986), and Heitjan and Rubin (1986).

2. AN ARTIFICIAL EXAMPLE ILLUSTRATING MULTIPLE IMPUTATION

Suppose we have taken a simple random sample of $n = 10$ units from a large population. The objective of the survey is to estimate \bar{Y} the mean of Y in the population. We know the mean value of a covariate X in the population, and the survey attempts to record both X and Y for each of the n units included in the sample.

Table 1 presents the observed values of (Y, X) for the ten units in the sample where the question marks indicate missing Y data due to nonresponse.

2.1 Multiply Imputing for the Missing Values

Suppose the missing values in Table 1 are to be multiply imputed using two values drawn under each of two models (i.e. two repetitions per model). In general, any number of models can be used with any number of repetitions within each model. Model 1 is an "ignorable" model for nonresponse; ignorable is defined precisely in Rubin (1976), but essentially it means

that a nonrespondent is only randomly different from a respondent with the same value of X . Model 2 is a nonignorable model and posits a systematic difference between respondents and nonrespondents with the same value of X . The repeated imputations under each model are based on a simple procedure closely related to the hot-deck, which can be improved upon but is useful to illustrate ideas.

For each nonrespondent, the two closest matches among the respondents are found, where the distance for matching is defined by the values of X . For the first nonrespondent, unit 2, the two closest matches are units 1 and 3, and for the second nonrespondent, unit 4, the closest matches are 3 and 5. The repeated imputations are created by drawing at random from the two closest matches. For the ignorable model, we simply impute the value Y provided by the matching respondent: the first two columns of Table 2 give the result. For the nonignorable model, we suppose that the nonresponse bias is such that a nonrespondent will tend to have a value of Y 20% higher than the matching respondent's value of Y : the last two columns of Table 2 give the result where the Y values have been rounded to the nearest integer. The repeated imputations within each model allow the user to draw a valid inference under that model. The use of two models, an ignorable one and a nonignorable one, allows the display of sensitivity of inference to assumptions about nonresponse. Generally such assumptions are untestable using the data at hand.

Table 1
Observed Data

Unit	Y	X
1	10	8
2	?	9
3	14	11
4	?	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25

Table 2
Multiple Imputations for Data of Table 1

	Model 1 Repetition		Model 2 Repetition	
	1	2	1	2
Unit 2	10	14	12	17
Unit 4	16	14	19	17

2.2 Analyzing the Resultant Multiply-Imputed Data Set

Each set of imputations, that is each column of Table 2, can be used with the incomplete data in Table 1 to create a completed data set. Since there are four sets of imputations, four completed data sets can be created; these are displayed in Tables 3 to 6. Each completed data set is analyzed just as if there had been no nonresponse.

Assume that with complete data, the ratio estimator $\bar{X}\bar{y}/\bar{x}$ would be used with associated variance SE^2 , where \bar{X} is the known mean of X in the population, say 12, \bar{y} and \bar{x} are the means of Y and X in the random sample of n units, and

$$SE^2 = \sum (Y_i - X_i\bar{y}/\bar{x})^2/[n(n - 1)]$$

Table 3
Complete Data Set 1 (Model 1, Rep. 1)
For Multiply Imputed Data Set of Tables 1 and 2

Unit	Y	X
1	10	8
2	10	9
3	14	11
4	16	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25
means	14.5	13

Table 4
Complete Data Set 2 (Model 1, Rep. 2)
For Multiply Imputed Data Set of Tables 1 and 2

Unit	Y	X
1	10	8
2	14	9
3	14	11
4	14	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25
means	14.7	13

Table 5
Complete Data Set 3 (Model 2, Rep. 1)
For Multiply Imputed Data Set of Tables 1 and 2

Unit	Y	X
1	10	8
2	12	9
3	14	11
4	19	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25
means	15	13

Table 6
Complete Data Set 4 (Model 2, Rep. 2)
For Multiply Imputed Data Set of Tables 1 and 2

Unit	Y	X
1	10	8
2	17	9
3	14	11
4	17	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25
means	15.3	13

Table 7
Ratio Estimates and Associated Variances of Estimates
for the Complete Data Sets of Tables 3-6

	Model 1 Repetition		Model 2 Repetition	
	1	2	1	2
Estimate	13.38	13.57	13.85	14.12
Variance	2.96	3.19	3.38	3.84

Table 8
Combined Estimates and Variances for the Multiply
Imputed Data Sets of Tables 1 and 2

	Model 1	Model 2
Estimate	13.48	13.98
Variance	3.10	3.66

where the sum is over the units in the sample. Table 7 presents the estimates and variances associated with each of the four completed data sets given in Tables 3-6.

The two answers obtained under the same model can be combined to obtain one inference for \bar{Y} under each model. The results are displayed in Table 8: the estimate is the average of the estimates and the variance associated with this estimate has two components: (i) the average within-imputation variance associated with the estimate and (ii) the between-imputation variance of the estimate. Thus, under Model 1, the estimate is $(13.38 + 13.57)/2 = 13.48$; the associated estimated average within variance is $(2.96 + 3.19)/2$, and the associated estimated between variance is $[(13.38 - 13.48)^2 + (13.57 - 13.48)^2]$. The estimated variances are combined as: (estimated total variance) = (estimated average within variance) + $(1 + M^{-1}) \times$ (estimated between variance), where the factor $(1 + M^{-1})$ multiplying the usual unbiased estimate of between variance is an adjustment for using a finite number of imputations. The associated 95% interval estimate for \bar{Y} is (10.0, 16.9) under Model 1 and (10.2, 17.7) under Model 2. In practice, better intervals can be formed by calculating degrees of freedom as a simple function of the variance components and using the 95% points appropriate to the corresponding t -distribution; when either M is large or the between variance component is small relative to the total variance (as in this artificial example), the degrees of freedom will be large and thus the normal 95% points will be used. Details are given in Section 3.

The essential feature to notice in this illustrative example is that only complete-data methods of analysis are needed. We merely have to perform the complete-data analysis that would have been used in the absence of nonresponse on each of the completed data sets created by the multiple imputations. The resultant answers under each model are then easily combined to give one inference under each model. Although not illustrated here, diagnostic analyses using complete-data techniques can be applied to each completed data set; Heitjan and Rubin (1986) provides several examples.

3. GENERAL PROCEDURES

The example in Section 2 illustrated methods for creating multiple imputations and analyzing the resultant multiply-imputed data set in a special case. We now outline the methods needed for general practice.

3.1 Proper Imputation Methods

Multiple imputations ideally should be drawn according to the following general scheme. For each model being considered, the M imputations of the missing values, Y_{mis} , are M repetitions from the posterior predictive distribution of Y_{mis} , each repetition being an independent drawing of the parameters and missing values under an appropriate Bayesian model for the posited response mechanism. In practice, implicit models such as illustrated

in Section 2 can often be used in place of explicit models. Both types of models are illustrated in Herzog and Rubin (1983), where repeated imputations are created using an explicit regression model and an implicit matching model, which is a modification of the Census Bureau's hot-deck.

Procedures that incorporate appropriate variability among the repetitions within a model are called *proper*, which is defined precisely in Rubin (1986a). The essential idea of proper imputation methods is to properly reflect sampling variability when creating repeated imputations under a model. For example, assume ignorable nonresponse so that respondents and nonrespondents with a common value of X have Y values only randomly different from each other. Even then, simply randomly drawing imputations for nonrespondents' from matching respondents' Y values ignores some sampling variability. This variability arises from the fact that the sampled respondents' Y values at X randomly differ from the population of Y values at X . Properly reflecting this variability leads to repeated imputation inferences that are valid under the posited response mechanism.

In the context of simple random samples and ignorable nonresponse, Rubin and Schenker (1986) study hot-deck imputation (i.e. simply randomly drawing imputed values from respondents), which is *not* proper, and a variety of proper imputation methods based on both explicit and implicit models, including a fully normal model, the Bayesian Bootstrap (Rubin, 1981), and an approximate Bayesian Bootstrap. The Approximate Bayesian Bootstrap (ABB) can be used to illustrate how an intuitive imputation method, such as the simple random hot-deck, can be modified to be proper.

3.2 Example of a Proper Imputation Method with Ignorable Nonresponse – The ABB

Consider a simple random sample of size n with n_R respondents and $n_{NR} = n - n_R$ nonrespondents. The ABB creates M ignorable repeated imputations as follows. For $\ell = 1, \dots, M$, create n possible values of Y by first drawing n values at random with replacement from the n_R observed values of Y , and second drawing the n_{NR} missing values of Y at random with replacement from those n values. The drawing of the n_{NR} missing values from a possible sample of n values rather than the observed sample of n_R values generates appropriate between imputation variability, at least in large samples, as shown by Rubin and Schenker (1986). The ABB approximates the Bayesian Bootstrap by using a scaled multinomial distribution to approximate a Dirichlet distribution.

3.3 Analysis – The Repeated Imputation Inference

The general methods for analyzing a multiply imputed data set implicitly assume proper imputation methods have been used to create the multiple imputations. As illustrated in Section 2, the repeated imputations within each model are analyzed as a collection to create one *repeated-imputation* inference as follows. Each data set completed by imputation is analyzed using the same complete-data method that would be used in the absence of nonresponse. More precisely, let $\hat{\Theta}_\ell, U_\ell, \ell = 1, \dots, M$ be M complete-data estimates and their associated variances for a parameter Θ , calculated from the M data sets completed by repeated imputations under one model for nonresponse. The final estimate of Θ is

$$\bar{\Theta}_M = \sum_{\ell=1}^M \hat{\Theta}_\ell / M.$$

The variability associated with this estimate has two components: the average within-imputation variance,

$$\bar{U}_M = \sum_{\ell=1}^M U_\ell / M,$$

and the between-imputation component,

$$B_M = \sum (\hat{\Theta}_i - \bar{\Theta}_M)^2 / (M-1)$$

where with vector Θ , $(\bullet)^2$ is replaced by $(\bullet)^T(\bullet)$. The total variability associated with $\bar{\Theta}_M$ is then

$$T_M = \bar{U}_M + (1 + M^{-1})B_M.$$

With scalar Θ , the reference distribution for interval estimates and significance tests is a t -distribution.

$$(\Theta - \bar{\Theta}_M) T_M^{-1/2} \sim t_v,$$

where the degrees of freedom,

$$v = (M - 1) \{ 1 + [(1 + M^{-1})B_M / \bar{U}_M]^{-1} \}^2$$

is based on a Satterthwaite approximation (Rubin and Schenker 1986 and Rubin 1986a). The within to between ratio \bar{U}_M / B_M estimates the population quantity $(1 - \gamma) / \gamma$, where γ is the fraction of information about Θ missing due to nonresponse. In the case of ignorable nonresponse with no covariates, γ equals the fraction of data values that are missing.

3.4 Significance Levels for Multicomponent Θ

For Θ with k components, significance levels for null values of Θ can be obtained from M repeated complete-data estimates, $\hat{\Theta}_i$, and variance-covariance matrices, U_i , using multivariate analogues of the previous expressions.

A simple procedure described in Li (1985) and Rubin (1986a) that works well for M large relative to k is to let the p -value for the null value Θ_0 of Θ be $\text{Prob} \{F_{k,v} > D_M\}$ where $F_{k,v}$ is an F random variable and $D_M = (\Theta_0 - \bar{\Theta}_M) T_M^{-1} (\Theta_0 - \bar{\Theta}_M)^T$ with v defined by generalizing B_M / \bar{U}_M to be the average diagonal element of $B_M \bar{U}_M^{-1}$, $\text{trace}(B_M \bar{U}_M^{-1}) / k$. Better procedures are described in Rubin (1986a). Less precise p -values can be obtained directly from M repeated complete-data significance levels; also see Rubin (1986a).

4. DISCUSSION

4.1 Frequency Evaluations

Although repeated imputation inferences are most directly motivated from the Bayesian perspective, they can be shown to possess good frequency properties. In fact, the definition of proper imputation methods means that in large samples infinite- M repeated imputation inferences will be valid. Since the finite- M adjustments are derived using approximations to Bayesian posterior distributions, however, deficiencies can arise with finite M . For example, the large sample relative efficiency of $\bar{\Theta}_M$ to $\bar{\Theta}_\infty$ that is, the efficiency of the finite- M repeated imputation estimator using proper imputation methods relative to the infinite- M estimator in units of standard errors is $(1 + \gamma / M)^{-1/2}$. Even for relatively large γ , modest values of M result in estimates $\bar{\Theta}_M$ that are nearly fully efficient.

4.2 Confidence Coverage

In large samples the confidence coverage of proper imputation methods using the t -reference distribution can be tabulated as a function of M , γ and the nominal level, $1 - \alpha$. Table 9 is from Rubin (1986a) and is also partially reported in Rubin and Schenker (1986) and Schenker (1985). Also included are results for single imputation, where the between component of variance is set to zero, since it cannot be estimated, and the reference distribution is the normal, since v cannot be estimated without B_M . Even in extreme cases, two or three repeated imputations yield nearly valid confidence coverages; this is in striking contrast to using only one imputation. Even worse coverages for single imputation would have been obtained using best prediction methods, such as “fill in the mean”.

Table 9

Coverage probabilities in % of interval estimates based on the t -reference distribution as a function of the number of proper repeated imputations, $M \geq 2$, the fraction of missing information, γ , and the nominal level, $1 - \alpha$. Also included for contrast are results based on single imputation $M = 1$, using the normal reference distribution with the between component of variability set to zero.

[illegible]

4.3 Significance Levels

Work on accurately obtaining significance levels is at an early stage of development. Table 10 is from Rubin (1986a) and is also partially reported in Li (1985). It indicates that if $M > k$ and γ is modest, accurate tests can be obtained using D_M . Better procedures are considered by Li (1985), Rubin (1986a) and in current thesis work by T.E. Raghunathan.

Table 10

Level in % of D_M with $F_{k, v}$ reference distribution as a function of: nominal level, α ; number of components being tested, k ; number of repeated proper imputations, M ; and fraction of missing information, γ .

k	M	$\gamma =$	$\alpha = 1\%$				$\alpha = 5\%$				$\alpha = 10\%$			
			.1	.2	.3	.5	.1	.2	.3	.5	.1	.2	.3	.5
2	2		1.0	1.2	1.6	2.5	4.9	5.3	5.9	7.5	9.9	10.3	11.0	12.9
	3		1.0	1.0	1.0	1.3	4.9	4.9	5.0	5.5	9.9	9.8	10.0	10.9
	5		1.0	1.0	1.1	1.2	5.0	5.0	5.1	5.6	10.0	10.0	10.2	10.9
	10		1.0	1.0	1.1	1.2	5.0	5.1	5.3	5.7	10.1	10.2	10.4	11.0
	25		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	10.0	9.9	9.9	10.0
	50		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	10.0	9.9	9.9	10.0
	100		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.1
3	2		1.0	1.1	1.3	1.7	5.1	5.3	5.6	6.3	10.3	10.6	11.1	12.0
	3		1.0	1.0	1.0	1.0	5.1	5.2	5.3	5.7	10.2	10.5	10.9	12.3
	5		1.0	1.0	1.1	1.3	5.0	5.2	5.4	6.2	10.1	10.3	10.8	12.2
	10		1.0	1.0	1.1	1.2	5.0	5.2	5.3	5.9	10.1	10.3	10.6	11.6
	25		1.0	1.0	1.1	1.2	5.0	5.1	5.2	5.6	10.1	10.2	10.4	10.9
	50		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.1	10.0	10.0	10.0	10.2
	100		1.0	1.0	1.0	1.0	5.0	5.0	5.1	5.1	10.0	10.0	10.1	10.2
5	2		0.9	0.8	0.8	0.9	5.1	4.8	4.5	4.0	10.5	10.4	10.1	9.2
	3		1.0	1.0	1.0	0.9	5.2	5.5	5.7	6.1	10.5	11.3	12.1	14.4
	5		1.1	1.1	1.2	1.4	5.2	5.6	6.1	7.7	10.4	11.1	12.2	15.4
	10		1.0	1.1	1.2	1.5	5.1	5.3	5.6	6.9	10.1	10.4	11.1	13.1
	25		1.0	1.0	1.1	1.3	5.0	5.2	5.3	6.0	10.1	10.3	10.6	11.5
	50		1.0	1.0	1.0	1.1	5.0	5.1	5.1	5.4	10.0	10.1	10.2	10.7
	100		1.0	1.0	1.0	1.1	5.0	5.0	5.1	5.2	10.0	10.1	10.1	10.4
10	2		0.8	0.5	0.3	0.1	5.1	4.0	2.9	1.5	10.8	10.1	8.5	5.4
	3		1.1	0.9	0.6	0.3	5.6	5.9	5.7	4.9	11.3	12.7	13.8	16.2
	5		1.1	1.2	1.3	1.4	5.4	6.3	7.4	11.0	10.7	12.4	14.8	22.7
	10		1.1	1.2	1.4	2.2	5.2	5.8	6.8	10.3	10.4	11.4	13.1	19.0
	25		1.0	1.1	1.2	1.6	5.0	5.2	5.6	7.1	10.0	10.4	11.0	13.4
	50		1.0	1.0	1.1	1.3	5.0	5.1	5.4	6.1	10.0	10.2	10.6	11.8
	100		1.0	1.0	1.1	1.2	5.0	5.2	5.3	5.8	10.1	10.2	10.5	11.3

5. CONCLUSION

In conclusion, multiple imputation is a very promising new tool for helping to handle nonresponse in surveys. Although much work remains to be done before it will become a commonplace method, many interesting theoretical and practical results suggest effort expended in its development will be well rewarded by important contributions to applied work.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation Grant Number SES-8311428. I wish to thank M.P. Singh and a referee for their helpful editorial comments on the earlier draft of this article.

REFERENCES

- HEITJAN, D.F., and RUBIN, D.B. (1986). Inference for coarse data using multiple imputation. *Proceedings of the 18th Symposium on the Interface of Computer Science and Statistics*.
- HERZOG, T.N., and RUBIN, D.B. (1983). Using multiple imputations to handle nonresponse in sample surveys. In *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliography*, New York: Academic Press, 209-245.
- LI, K.H. (1985). *Hypothesis Testing in Multiple Imputation - with Emphasis on Mixed-up Frequencies in Contingency Tables*. Ph.D. Thesis, Department of Statistics, University of Chicago.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- RUBIN, D.B. (1977). The design of a general and flexible system for handling nonresponse in sample surveys. Unpublished paper prepared for the U.S. Social Security Administration.
- RUBIN, D.B. (1978). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20-34. Also in *Imputation and Editing of Faulty or Missing Survey Data*, U.S. Dept. of Commerce, 1-23.
- RUBIN, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. *Proceedings of the 1979 Meetings of the ISI-IASS, Manila*.
- RUBIN, D.B. (1980). *Handling Nonresponse in Sample Surveys by Multiple Imputations*. U.S. Dept. of Commerce, Bureau of the Census Monograph.
- RUBIN, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.
- RUBIN, D.B. (1986a). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- RUBIN, D.B. (1986b). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 87-94.
- RUBIN, D.B., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- SCHENKER, N. (1985). *Multiple Imputation for Interval Estimation from Surveys with Ignorable Nonresponse*. Ph.D Thesis, Department of Statistics, University of Chicago.

Imputation Options in a Generalized Edit and Imputation System

P. GILES and C. PATRICK¹

ABSTRACT

Statistics Canada has undertaken a project to develop a generalized edit and imputation system, the intent of which is to meet the processing requirements of most of its surveys. The various approaches to imputation for item non-response, which have been proposed, will be discussed. Important issues related to the implementation of these proposals into a generalized setting will also be addressed.

KEY WORDS: Modularity; Prototyping; Donor imputation; Regression models.

1. GENERALIZED SYSTEMS

Due to resource constraints imposed on surveys in recent years, especially in the area of development, the idea of generalized software has received considerable support. By generalized software, it is meant a set of computer programs, tied together into one system, which allows the user to select a suitable approach to the problem, from among several alternatives. For example, a user has a data file from which a sample of records is to be selected. A generalized sample selection system would offer the user the choice of various sampling schemes such as simple random or unequal probability sampling (with or without replacement), systematic, stratified, or cluster sampling.

A genuinely generalized system is, almost by definition, a complex object. The concept of modularity is an important device for the reduction of complexity, by allowing the overall task to be split into a number of simpler sub-tasks. Each of the sub-tasks, or functions, is performed sequentially. The user is offered several alternatives for each sub-task. Therefore, not only is the overall task able to be split into smaller, more manageable components, but also each sub-task can be performed in more than one way.

Figure 1 demonstrates how the edit and imputation task can be split into three sub-tasks. These three sub-tasks are editing, identification of fields to impute, and imputation. Each of the boxes, or modules, in a row employ different approaches to that particular sub-task. For example, C1 could employ some type of donor imputation, C2 could employ the imputation of a mean value, and so on. The user would select one of the modules from each of rows A, B, and C.

It should be noted that this representation of a generalized system for edit and imputation is not the only possibility. In fact, the actual proposal for a developmental project actually contains five sub-tasks, as opposed to the three exemplified here. This representation is given only for simplicity.

Each sub-task, or row in the example, would be a clearly defined function. The input files required, and the output files created, must have prespecified formats. This allows the user to concentrate on the choice of modules in each row, knowing that the system can handle the "housekeeping". (This refers to file handling and other mundane details about which the user would prefer not to worry.) Even though the system may accept all possible combinations of choices of modules, some combinations may not be desirable or even logically valid. It is usually the responsibility of the user to ensure that the pieces fit together.

¹ Philip Giles and Charles Patrick, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

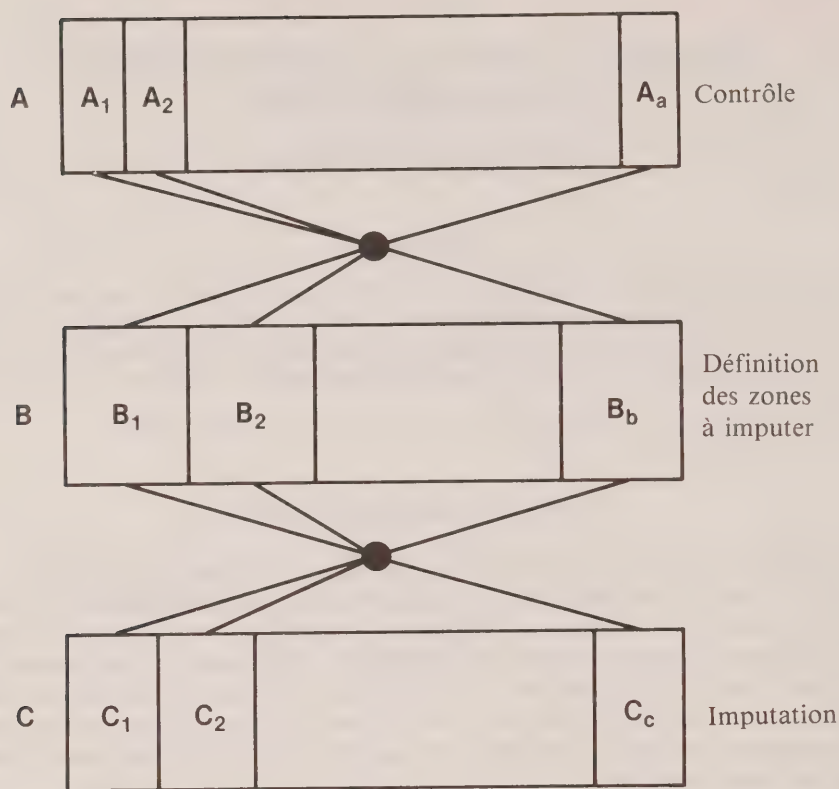


Figure 1. Generalized System Example – Edit and Imputation

A modular approach to the development of a processing system has an important consequence. From a certain point of view, the system is always “under development”, since additional modules embodying new approaches and enhancements to “old” modules, can always, and in principle should, be added. This open-endedness also means that the very important concept of prototyping can be easily accommodated. Prototyping is an approach wherein a subset of modules are developed initially. The system would then be available to some of the users. Subsequently, additional modules are developed to meet the requirements of additional users. Thus, the key advantage of prototyping and modularity is that piecemeal improvements to the system are deliberately anticipated and more easily accomplished. A minimal, but imperative, requirement of such an approach is that a framework (as shown in Figure 1) and a host environment (format of data files and programming language) must be carefully defined and specified very early in the overall developmental process.

In addition to the foregoing developmental advantages, others may be gained after the system is in place. The user has considerable flexibility in choosing the path to proceed. If several alternatives seem equally viable, one can use historical data to choose among them, by testing the various alternatives prior to data collection. This can be accomplished without an undue expenditure of effort. Once the generalized system is developed there is a reduction in resource requirements for each of its users, with a corresponding reduction in elapsed time to implementation.

There are some disadvantages to following a generalized route. The utilization of generalized software in a production environment may be less efficient than the corresponding custom-designed system. The initial resource requirement will be higher for a generalized system as compared to a customized system. However, this higher cost must be assessed against the

substantially higher costs of repeated custom-designed implementations. Nor is it reasonable to expect a generalized system to satisfy every specific requirement. In this situation, the user has two options. The first option is to develop a user-written module. This would not require the same degree of effort as a complete customization. However, if this occurs frequently, the purpose of the generalized system is defeated. The second option is for the user to modify the specifications in order to fit the generalized system mold. If the system has been well-designed, any required compromise should not result in a serious deterioration of data quality. It should also be recognized that compromises to the original specifications are usually and frequently required during the development of a customized system.

2. BACKGROUND TO IMPUTATION

The term "imputation", in this document, refers to a certain class of procedures for handling non-response. The input is a data captured file. The imputation procedure creates a file with individually "clean" records; a "clean" record being one which has no missing values and which satisfies all the specified edits. In order to create a clean record, a value must be estimated for each missing value.

The edits, specified by the user, are logical constraints on the values that each variable can assume. The set of edits, as a whole, define the acceptance region for the data. For categorical data, an edit is specified as a set of combinations of acceptable data values. The acceptance region can be represented as a set of lattice points in N -space. For numerical data, an edit is a linear equality or inequality. The requirement of linearity is not unduly restrictive, since a non-linear edit can be made linear by either algebraic manipulation or by adding supplementary variables, which are suitably defined non-linear functions of survey variables. The acceptance region for numerical data is a set of convex regions in N -space. The reason that there may be more than one convex region is that conditional edits are possible. Conditional edits are edits which pertain to only a subset of records. For example, the edits which are relevant to a particular record may be very different, depending on whether the variable Sex is recorded as Male or Female.

If one or more edits fail for a particular record, it may not be obvious which variable(s) is/are in error, and, by implication, to be imputed. For example, a failed edit is $A + B \leq C$. The data record under consideration has data values $A = 10$, $B = 5$, $C = 12$. There are seven combinations of variables to change which would result in a clean record. These are A , B , C , $A \& B$, $A \& C$, $B \& C$, and, $A \& B \& C$. Without any other information or decision rule, each of these choices is equally valid. The problem of how to decide which variable(s) to impute will not be discussed in this document. It will be assumed that, for each record, the variable(s) to impute have been identified. No distinction is made between variables to impute due to missing values and variables to impute due to edit failures.

3. PROPOSED IMPUTATION TECHNIQUES

This section is comprised of four sub-sections, which define all the proposed imputation techniques. These are Deterministic Imputation, Donor Imputation, Regression Models, and Other Imputation Estimators. The use of regression models and the section on other estimators is restricted to numerical data. The other two sub-sections apply both to numerical and categorical data.

Almost all imputation techniques can be formulated in a prediction framework, described by Rubin (1976), as follows. A joint distribution, $f(X_1, \dots, X_N)$, summarizing the

statistical behavior of the population of complete records is specified. This can be done whether the individual variables are quantitative or qualitative. Without loss of generality, for a record i which requires imputation, the N variables can be partitioned into X_1, \dots, X_{m_i} , which require imputation, and X_{m_i+1}, \dots, X_N , which do not require imputation. A conditional distribution $f(X_1, \dots, X_{m_i} | x_{m_i+1}, \dots, x_N)$ can be derived. Imputed values, y_1, \dots, y_{m_i} , are chosen for X_1, \dots, X_{m_i} from the set.

$$\{y_1, \dots, y_{m_i} : f(y_1, \dots, y_{m_i} | x_{m_i+1}, \dots, x_N) > 0\}$$

Various selection mechanisms can be employed. However, as stated above, some of these are relevant only to certain types of data variables.

It should be noted that there is nothing new or radically different in these proposals. They are based on work done previously, both in Statistics Canada and outside. The discussion on donor imputation is based on Fellegi and Holt (1976). The model-based approach to determining a value to impute is discussed by Little (1982). Other related papers of interest are Sande (1976), Kalton and Kasprzyk (1982), and Kalton and Kish (1981).

3.1 Deterministic Imputation

The first type of imputation is called deterministic imputation. This occurs when only one value can satisfy the edits. If more than one variable is to be imputed for a particular record, a deterministic solution may be possible for some, or all, variables. The check for determinacy should be done before proceeding to other imputation procedures.

Deterministic imputation may arise in very simple, and easily detectable situations. For example, suppose that there is an edit $A + B = 10$. The record under consideration requires A to be imputed and B has value 6. Obviously, $A = 4$ is the only value which will satisfy the edit. Another example demonstrates this for categorical variables. Suppose an edit is stated as "If the relationship to the household reference person is wife, then sex must be female." If the reference record has "wife" as the value of "relationship to the household reference person", and the variable "Sex" requires imputation, then the only valid imputed value is Sex = Female.

However, a typical survey situation will have several edits, rather than just one. This may mean that an existing deterministic solution may not be apparent. The procedure for checking for deterministic imputation is to find the reduced acceptance region defined by the active edits and the "good" data values. The active edits are defined as the subset of edits in which the variable(s) to be imputed are participant. This can also be expressed in the notation of the prediction framework given at the beginning of Section 3. The conditional distribution $f(X_1, \dots, X_{m_i} | x_{m_i+1}, \dots, x_N)$ will specify a unique value for some or all of the variables X_1, \dots, X_{m_i} .

An example serves to illustrate the procedure for identifying deterministic imputation. Note that while the example is written with numerical variables, an analogous situation exists for categorical variables.

There are three edits:

$$X + Y \leq 16,$$

$$Y + Z \leq 4,$$

$$X - 3Z \leq 8.$$

The reference record has values

$$X = 11 \text{ and } Y = 3.$$

The variable Z is to be imputed.

It is not apparent whether or not a determinancy exists. This first step is to consider all active edits. In the example, there are two edits which contain the variable Z .

$$Y + Z \leq 4,$$

$$X - 3Z \leq 8.$$

Next, the known values of X and Y are inserted into these edits, and the reduced acceptance region is determined.

$$3 + Z \leq 4,$$

$$11 - 3Z \leq 8.$$

Solving these inequalities gives the following solution.

$$Z \leq 1,$$

$$Z \geq 1.$$

It is now obvious that $Z = 1$ is the only possible valid imputed value.

In most “real-life” situations, the incidence of deterministic imputation should be low. The contrary would indicate that the edits are more restrictive than necessary or desirable, and should lead to a re-examination of the edit specifications. However, in the sense that it reduces the imputation problem, deterministic imputation is a useful first step.

3.2 Donor Imputation

Donor imputation is a method which pairs each record requiring imputation, the candidate record, with one record from a defined donor population. In order to determine the value to impute, one approach is to directly copy the value from the donor record onto the candidate record. For numerical variables, if suitable auxiliary information is available, more complex methods may be used to determine the value to be imputed. Further discussion on imputation estimators for donor imputation is given in Section 3.3.

Usually, the donor population is defined as all records in the current survey which have no variables to be imputed. Referring to the prediction framework described at the beginning of Section 3, then this situation implies that $f(X_1, \dots, X_N)$ is the empirical probability function. However, other approaches to defining the donor population are possible. For the remainder of the discussion on donor imputation, it will simply be assumed that a donor population has been defined.

Donor-candidate pairs are formed using matching variables. Matching variables are defined as variables which do not require imputation on the candidate record and are “highly correlated” with the variable(s) requiring imputation. Preferably, the matching variables should also have “low correlation” with each other. Two matching variables with “high correlation” would have the same discriminatory power as one alone, but would have the effect of doubling the weight given to one alone.

For categorical variables, a donor record is chosen, using some random process, from amongst potential donor records having the same values for the matching variables to those for the candidate record. Since numerical variables can assume many more values than categorical variables, it is very unlikely that an exact match on matching variables would be possible. Therefore, for numerical data, a distance function is used to define similarity. This distance function is a function of the matching variables on the candidate and potential donor records. The chosen donor is the record with minimum distance from the candidate record. Usually, the matching variables are transformed for the purpose of distance calculations in order to remove the effect of scale in which the variable is recorded. For example, it would be quite worrisome to the user if the formation of the donor-candidate pairs was dependent on whether a length variable was recorded in metres or feet. The proposed transformations and distance functions are discussed below.

The matching variables to be used can be a user input, or determined by an automated procedure. Usually, due to time considerations, all decisions must be made prior to data collection. Therefore, if the determination of matching variables is a user input, the user must specify the matching variables for each pattern of variables to be imputed. If there are N variables on the file, the user must make $(2^{**}N) - 2$ input specifications. Obviously, the value of N does not have to be very large in order for this approach to become unmanageable. In order to reduce this number, the matching variables may be specified by stratum. All candidate records in a particular stratum would use the same matching variables. In this situation, it is possible (depending on how careful the user is in specifying the matching variables) that a particular candidate record may have a matching variable which requires imputation. All in all, the user who inputs the matching variable specifications, is warned that this decision may result in a large increase in the work required.

One possible approach for automatically determining the matching variables is proposed. This procedure can be used, analogously, for both categorical and numerical data. Basically, the procedure is as follows. At a minimum, the set of matching variables must contain the variables sharing in the edit rules with the variables to be imputed. As defined earlier, these are the active edits. This approach seems intuitively reasonable, since it is desirable that the matching variables be correlated with the variable(s) to be imputed. The variables in the active edits constrain the range of possible values to be imputed. This implies a type of dependence, or correlation structure.

The use of this matching procedure, together with direct transcription, has one important consequence for categorical variables. All imputed values are guaranteed to pass the edits. This is very important as it is required in order to create a clean record. Without this guarantee, the user must re-edit the records, and possibly adopt a secondary imputation procedure. For numerical data, similarity as defined by a distance function does not guarantee this outcome. However, the closer the distance between the donor and candidate record is to zero, the greater the probability that the imputed values will satisfy the edits.

The determination of matching variables using this automated procedure can be illustrated by an example.

There are five edits:

- I. $A + B \leq \alpha_1,$
- II. $B - E \leq \alpha_2,$
- III. $C + 2D + 3E \leq \alpha_3,$
- IV. $A + C + D \leq \alpha_4,$
- V. $A - 2B + C \leq \alpha_5.$

There are five survey variables A, B, C, D, E and $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ are known scalars. The candidate record under consideration has variable B only to be imputed.

The first step is to identify the active edits. In this example, there are three active edits. These are edits I, II, and V.

The second step is to determine the active variables. The active variables are defined as all variables which are contained in at least one of the active edits. In the example, there are four active variables: A, B, C, E . Note that, by definition, the active variables contain all variables to be imputed.

The third step is to determine the matching variables, as those active variables which do not require imputation. For this example, the matching variable are A, C, E .

In addition to the determination of matching variables, donor imputation for numerical data requires the choice of a data transformation and the choice of a distance function.

Two types of data transforms are proposed. For both of these, each variable is to be transformed independently. The two proposed transformations are a rank value transform and a location-scale transform.

For the rank value transform, the values for each variable are sorted. Then, the rank values are divided by a suitable constant such that all values are in the range from zero to one. The transformed values are distributed uniformly in that range.

The location-scale transform is of the form,

$$y^T = \frac{1}{b} (y - a),$$

where y^T is the transformed value,

y is the original data value,

a, b are user-specified parameters.

Two popular choices for these constants are, one, that a be the sample mean and b be the sample standard deviation, and, two, that a be the sample minimum and b be the range of values in the sample. Other options may be possible.

In choosing a data transform, there are robustness and outlier considerations. The rank value transform is very robust against changes in data values, and pulls outliers closer to the other data values. This may or may not be desirable. There are no bounds on the transformed values, using the location-scale transform with the mean and standard deviation. These parameters are also sensitive to outliers. The choice of the minimum value and range would restrict the transformed values between zero and one. However, these are very sensitive to extreme values. One very large value could cause all of the transformed values, except one, to be virtually zero.

In considering the choice of distance function, a family of distance functions are proposed. These are the weighted \mathcal{L}^p norms, where p is a user-specified constant. The general form of these functions is

$$D(X, Y) = \left[\sum_{k=1}^r w_k |x_k - y_k|^p \right]^{1/p},$$

where x_k, y_k are the r matching variables on the two records,

w_k are user-specified weights,

p is a user-specified constant.

The weights are used if one wishes some of the matching variables to contribute more to the distance calculation than others. The default values are for all weights to be set to one.

Three particular choices of a value for p are of special interest, $p = 1$, $p = 2$, and $p = \infty$. For $p = 1$, this function calculates the city block distance. For $p = 2$, the Euclidean distance is calculated. The limiting case of this function, when $p = \infty$, yields the minimax distance. For this choice of p , the function is written as

$$D(X, Y) = \text{Max}_{1 \leq k \leq r} [w_k |x_k - y_k|].$$

One final point to be discussed about donor imputation is the concept of a "penalty" for donor usage. This penalty would reduce the number of times that a particular donor record is used. For donor imputation of categorical data, a donor record is selected from the donor population without replacement. This strategy has to be modified slightly if the size of the candidate population is greater than the size of the donor population.

For numerical data, the distance function is modified by increasing the distance calculation according to the number of times a particular donor is used. One possible approach is to use $D'(X, Y)$ to calculate distances, where

$$D'(X, Y) = D(X, Y) \times (1 + ud),$$

where u is the "penalty" imposed by the user,

d is the number of times that donor record has been chosen.

An implication of the imposition of a penalty on the distance function, is that the choice of a donor record for each candidate record is now dependent on the order of the candidate records.

3.3 Regression Models

This section discusses imputation estimators which result from the use of regression models. For this discussion, only two models are used. These are:

$$\text{MODEL I : } y_i = \alpha + \epsilon_i, \quad \text{Var}(\epsilon_i) = \sigma^2,$$

$$\text{MODEL II: } y_i = \beta x_i + \epsilon_i, \quad \text{Var}(\epsilon_i) = \sigma^2 x_i.$$

Note that these models are special cases of the more general formulation of regression models, which has the form

$$\underline{y} = \underline{X}\underline{\beta} + \underline{\epsilon},$$

$$\text{where } E(\underline{\epsilon}) = \underline{0}, \quad V(\underline{\epsilon}) = \underline{V}$$

Model II is used when auxiliary data is available. Otherwise Model I is used. Both models have one parameter to be estimated. Using least-squares, the parameter estimates are:

$$\hat{\alpha} = \bar{y},$$

$$\hat{\beta} = \frac{\bar{y}}{\bar{x}}.$$

Before stating the various proposed estimators, some notation will be introduced.

- Let t be the subscript for time t , the present survey,
- y_{it} be the variable under study for unit i and time t ; this is the value to be imputed for candidate records,
- x_{it} be the auxiliary variable (correlated with Y) for unit i and time t ,
- R be the subscript for all non-respondents at time t (i.e., y_{it} is known),
- NR be the subscript for all non-respondents at time t (i.e., y_{it} is to be imputed),
- C, D be superscripts which denote either a candidate or donor record, whenever the distinction is required.

Several explanatory notes are required along with the notation. First, R and NR are as defined in the current survey, regardless of the reporting history of each record. Second, the values for the variables $y_{i(t-1)}, x_{it}, x_{i(t-1)}$ may themselves have been imputed. The only restriction is that they are not missing. Third, the notation does not include the concept of imputation classes. Imputation classes are essentially post-strata, in that they define sets of records which are judged homogeneous within, and heterogeneous between groups. However, both the notation and the imputation estimators are readily extendible to include imputation classes.

Thus, estimators can be classified according to:

- (i) the choice of model, I or II,
- (ii) the imputation group, and,
- (iii) the variables in the regression used to estimate the parameter.

The data on the records in the specified imputation group are precisely the data used to estimate the parameter(s) in the model. This concept allows considerable flexibility. For example, it could allow the preclusion of outliers from the calculation of the parameter estimate. After the parameter is estimated, it is used for prediction purposes to determine the imputed value. According to the notation, Y_i is always the variable predicted.

Based on the two models, eight imputation estimators are proposed. Even though there are eight proposed estimators, this list can be augmented in the future. These additional estimators could be derived, for example, by choosing other models, possibly incorporating more variables.

Scanning the list of eight, one can see that these are the familiar imputation estimators that have been used traditionally.

Estimator 1: The value from the previous survey for the same unit is imputed. $y_{i(t-1)}$

Estimator 2: The mean value from the previous survey is imputed. $\bar{y}_{(t-1)}$

Estimator 3: The mean value of all respondents to the current survey is imputed. \bar{y}_{tR}

Estimator 4: The value is copied directly from the donor record to the candidate record, y_{it}^D

Estimator 5: A ratio estimate, using values from the current survey is imputed.

$$\frac{\bar{y}_{tR}}{\bar{x}_{tR}} x_{it}$$

Estimator 6: A ratio estimate, based on values on the donor and candidate records is imputed.

$$\frac{y_{it}^D}{x_{it}^D} x_{it}$$

Estimator 7: The value from the previous survey for the same unit, with a trend adjustment calculated from an auxiliary variable, is imputed.

$$\frac{y_{i(t-1)}}{x_{i(t-1)}} x_{it}$$

Estimator 8: The value from the previous survey for the same unit, with a trend adjustment calculated from the change in reported values to variable Y , is imputed.

$$\frac{\bar{y}_{tR}}{\bar{y}_{(t-1)R}} y_{i(t-1)}$$

It is interesting to contrast the difference in estimators when one fixes all classification items but one. For example, the difference between estimators one and two is due only to the difference in choice of imputation group, as is also the case for estimators three and four, and, estimators five and six. The difference between estimators one and seven is due only to the choice of model. The same is true for estimators three and five, and, estimators four and six. It should also be noted that estimators four and six are those used in donor imputation, which were discussed in Section 3.2.

3.4 Other Imputation Estimators

The choice of imputation techniques is dependent upon the assumptions made by the user about the non-responding population. When using donor imputation, one assumes that there are some respondents which are similar to each non-respondent. If one imputes the mean from the current survey, the assumption is that the mean value of the respondents is the same as the mean value of the non-respondents. Similarly, one can go through all the estimators and list the implied assumptions. The first estimator proposed in this section tries to ease the somewhat restrictive (and usually untrue) assumptions required in the previous section. It pays for this by being more complex. It is called the chain-link estimator, given by Madow and Madow (1978).

The derivation of this estimator is described. First, by assuming that the rate of change (trend) of the non-responding and responding populations are the same as observed in the previous survey, the population mean of the variable Y for the non-responding population in the current survey is estimated.

$$\bar{y}_{NRt} = \frac{\bar{y}_{NR(t-1)}}{\bar{y}_{R(t-1)}} \bar{y}_{Rt}$$

One then determines the imputed value according to the auxiliary variable.

$$\begin{aligned} y_{it} &= \frac{\bar{y}_{NRt}}{\bar{x}_{NRt}} x_{it} \\ &= \frac{\bar{y}_{NR(t-1)}}{\bar{x}_{NRt}} \frac{\bar{y}_{Rt}}{\bar{y}_{R(t-1)}} x_{it} \end{aligned}$$

Note that this amounts to a more complex application of the Regression Model approach discussed in Section 3.3. First, temporarily impute $y_{it} = \bar{y}_{NRt}$, as given above. Then, use Model II, and define the imputation group as being all non-responding records to the present survey for variable Y . The response variable is Y_t . The regressor variable is X_t . The resulting estimator is as given above.

The second estimator proposed in this section can be used when one has data on variable Y for several previous surveys. It does not use auxiliary variables, or data from other records. The behavior of each non-respondent is considered independently of others. This method is called exponential smoothing. It is a standard econometric forecasting technique. There is one user-specified parameter. It allows the flexibility of changing the relative contribution of the various data values. Algebraically, the estimator is given by

$$y_{it} = \frac{1-A}{1-A^t} \sum_{r=0}^{t-1} A^r y_{i(t-r-1)},$$

where $0 < A < 1$, is prespecified.

The closer A is to zero, the more weight is given to recent data. If $t = 1$, this reduces to imputing the value for the previous survey.

4. PAST WORK IN STATISTICS CANADA

Statistics Canada has made efforts in the past to develop a generalized edit and imputation system. Two of these will be highlighted, as they form the basis for the current proposal. These are the CAN-EDIT system and the Numerical Edit and Imputation System (NEIS).

4.1 CAN-EDIT

CAN-EDIT is itself, not a completely generalized system. However, the methodology that it employed is. The system is based on the work by Fellegi and Holt (1976) on imputation for categorical data. It was developed for processing the 1976 and 1981 Canadian Censuses of Population and Housing.

CAN-EDIT adopted a donor imputation approach. The matching variables were determined automatically, using the procedure described in Section 3.2. The CAN-EDIT system employed what it called primary and secondary imputation. If a candidate record could not be imputed in primary imputation, it was sent to secondary imputation.

In primary imputation, all imputed values are taken from the same donor. The matching variables were determined based on all variables to be imputed. A record would fail primary imputation if no donor record had identical values on the matching variables.

In secondary imputation, each of the variables to be imputed are treated independently and sequentially. The procedure for determining the matching variables is the same. However, by considering only one variable at a time, the number of matching variables will, in general, be less than under primary imputation. (There cannot be more, but the number may be the same). This implies that the potential donor population is larger. There are a few disadvantages to secondary imputation, as compared to primary imputation. First, it is possible to choose, as a matching variable, a variable which is to be imputed. There is no value to match on. Second, this approach does not make use of the joint distributions of the variables. The imputed values for two variables may satisfy the edits, each may be a very valid value, but which may occur in the population in combination only rarely.

4.2 Numerical Edit and Imputation System (NEIS)

The NEIS is a first prototype of a generalized E&I system for numerical data. It was written as a set of modules in the PSTAT statistical package. Subsequent prototypes have never

been developed. This system was developed by Gordon Sande (1979). It is felt that the methodology is very sound, and should be incorporated in a new system. However, PSTAT may no longer be a suitable software environment. The NEIS was used, in a production environment, by the 1981 Farm Energy Use Survey. The methodology was employed in the development of the 1981 Census of Agriculture processing system.

The NEIS, similar to CAN-EDIT, used a donor imputation approach with matching variables determined automatically using the procedure described in Section 3.2. However, as explained in that section, the determination of matching variables in this fashion for numerical data will not always result in the imputation procedure producing a clean record. The strategy adopted to reduce this problem is to select the closest r donors. If the closest donor does not impute values which satisfy the edits, then the next closest donor is considered, and so on.

The NEIS gave the user no choice of transformation or distance function. It used the rank value transformation and the weighted L^∞ norm for distance calculations.

5. CONCLUSION

The proposals presented would allow considerable choice to a user of a generalized edit and imputation system. As mentioned, it does not close the door on additional approaches. However, it is felt that a system which is developed with these components would be suitable for a large number of users. It has been the experience of the authors that the ultimate power and usefulness of such a system is not apparent until one starts to use it. As testing proceeds, it becomes clear that there are more capabilities and extensions than first appear.

REFERENCES

- FELLEGI, I.P., and, HOLT, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.
- KALTON, G., and, KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 22-31.
- KALTON, G., and KISH, L. (1981). Two efficient random imputation procedures. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 146-151.
- LITTLE, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- MADOW, L.H., and MADOW, W.G. (1978). On link relative estimators. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 534-539.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SANDE, G. (1976). Searching for numerically matched records. Technical Report, Business Survey Methods Division, Statistics Canada.
- SANDE, G. (1979). *The Numerical Edit and Imputation Subsystem for PSTAT — A User's Guide*. Research and General Systems Subdivision, Statistics Canada.

The Maximum Likelihood Method for Non-Response in Sample Surveys

M.S. SRIVASTAVA and E.M. CARTER¹

ABSTRACT

The analysis of survey data becomes difficult in the presence of incomplete responses. By the use of the maximum likelihood method, estimators for the parameters of interest and test statistics can be generated. In this paper the maximum likelihood estimators are given for the case where the data is considered missing at random. A method for imputing the missing values is considered along with the problem of estimating the change points in the mean. Possible extensions of the results to structured covariances and to non-randomly incomplete data are also proposed.

KEY WORDS: Incomplete response; Missing at random; Maximum likelihood method; Imputation.

1. INTRODUCTION

Examples of non-response in sample surveys are in abundance. Various attempts with varying degrees of success have been made in the literature to solve this problem. The success of a particular procedure is dependent on the complexity of the problem. For example, when the data is not missing at random, the problem is far from being solved. The recent attempts by Heckman (1976) and Greenlees *et al.* (1982) among others, are highly sensitive to model misspecification. Similarly the hot-deck method has been severely criticized in the literature. However, when the sample size is large, the hot-deck method and a carefully designed regression method yield similar results in imputing the non-response income in Current Population Survey (CPS). See David, Little, Samuhel and Triest (1986).

The regression method is based on the assumption that the non-response is random, but unlike the hot-deck method does not require complete information from a previous census, which in a majority of cases is non-existent. Thus it appears that a carefully designed regression method may be of great help.

In this paper, the situation when the non-response is random is considered. Random non-response arises naturally in many situations. For example, in successive sampling, the sampling starts with a certain number of people from whom certain observations are obtained for a period of time. At the end of this period, some people are dropped from the survey and new people are added. The survey continues in this manner until completion. Examples of this nature are considered by Woolson, Leeper and Clarke (1978) and Woolson and Leeper (1980).

Even when the non-response is not random, the non-random nature of the incomplete data may be accounted for, by using a sufficient number of explanatory variables in the regression model and employing some of the techniques used in the hot-deck method as was done in David *et al.* (1986) for a univariate model. For example, in Section 2.5 a method for imputing the missing values is given.

¹ M.S. Srivastava, Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1, and E.M. Carter, Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, Canada N1G 2W1.

In the course of developing these results, a method will be derived for checking if there have been any changes over time in the response patterns. The models used can also be modified to include error variance-covariance matrices that are structured by the imposition of a time series to the reponse variables. In this paper it is assumed that the data are normally distributed from a simple random sampling scheme and that the data are missing at random. If the normality assumptions is dropped then the estimators can no longer be considered maximum likelihood estimators but may still be considered as good heuristic estimators.

In the next section, the form of the model will be described for the one sample problem.

2. THE ONE SAMPLE PROBLEM

2.1 The Model

The bivariate incomplete data problem is considered first to introduce the general procedure that follows. Let $y = (y_1, y_2)'$ be a bivariate random vector with mean vector $\underline{\mu}$ and covariance matrix Σ . Without loss of generality, the missing data in the bivariate situation can be described as follows:

$$\begin{aligned} &y_{11}, \dots, y_{1n_1}, y_{1,n_1+1}, \dots, y_{1,n_1+n_2}, \text{-----} \\ &y_{21}, \quad, y_{2n_1}, \text{-----} y_{2,n_1+n_2+1}, \dots, y_{2,n_1+n_2+n_3} \end{aligned} \tag{1}$$

That is, there are n_1 pairs of observations, n_2 observations on y_1 with the corresponding observation on y_2 missing, and n_3 observations on y_2 with the corresponding observation on y_1 missing. Thus $N = n_1 + n_2 + n_3$ observations are grouped into three subsets. If the complete data set were to be represented as $\underline{y}_1, \dots, \underline{y}_N$, then the actual observed responses can be defined as

$$\begin{aligned} \underline{z}_{1j} &= B_1 \underline{y}_j = \underline{y}_j, \text{ for } j = 1, \dots, n_1, \\ \underline{z}_{2j} &= B_2 \underline{y}_j = y_{1j}, \text{ for } j = n_1 + 1, \dots, n_1 + n_2, \\ \text{and} \\ \underline{z}_{3j} &= B_3 \underline{y}_j = y_{2j}, \text{ for } j = n_1 + n_2 + 1, \dots, n_1 + n_2 + n_3, \end{aligned}$$

where $B_1 = I_2$, the identity matrix, $B_2 = (1 \ 0)$ and $B_3 = (0 \ 1)$.

For the general multivariate one sample problem, there will be K subsets of the data containing n_1, \dots, n_K observations. Note that the maximum number of groups is $2^p - 1$. Also the total sample size is $N = n_1 + \dots + n_K$. If the k -th subset contains p_k characteristics i_1, \dots, i_{p_k} , then the matrix B_k would be a $p_k \times p$ matrix with a one in the (s, i_s) position for $s = 1, \dots, p_k$ and zero elsewhere. With this notation the observed vectors of responses can be written as:

$$\underline{z}_{kj} = B_k \underline{y}_{kj}, \ j = 1, \dots, n_k, \ k = 1, \dots, K.$$

Hence,

$$E(\underline{z}_{kj}) = B_k \underline{\mu},$$

and

$$\text{cov}(\underline{z}_{kj}) = B_k \Sigma B_k', \ j = 1, \dots, n_k \text{ and } k = 1, \dots, K.$$

Example 1: (Data)

Wei and Lachin (1984) give the cholesterol levels for a treatment group studied at times 0, 6, 12, 20 and 24 months. For reasons not pertaining to the response variable, certain observations were incomplete. The data can be grouped into $K = 8$ subsets. For the first group of complete data the sample mean and covariance matrix, based on 36 observations, were:

$$\underline{\bar{z}}_1 = \begin{bmatrix} 226.6 \\ 249.6 \\ 252.6 \\ 253.1 \\ 256.7 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 1964 & 1301 & 1151 & 960 & 1008 \\ 1301 & 1715 & 1109 & 1023 & 1199 \\ 1151 & 1109 & 1554 & 697 & 1266 \\ 960 & 1023 & 697 & 1148 & 667 \\ 1008 & 1199 & 1266 & 667 & 2546 \end{bmatrix}.$$

The data for each of the other subsets is given in Table 1 with the imputed values in parenthesis.

The matrices that define the model for the observed values are:

$$B_1 = I_5, \quad B_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad B_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$
$$B_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad B_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad B_6 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$
$$B_7 = (1 \ 0 \ 0 \ 0 \ 0), \quad B_8 = (0 \ 1 \ 0 \ 0 \ 0).$$

Now that the model is defined, estimation of the parameters and the imputation of the missing data can be performed.

2.2 Estimation of the Population Mean Vector and Covariance Matrix.

For each of the K subsets define the sample mean as

$$\underline{\bar{z}}_k = (n_k)^{-1} \sum_{j=1}^{n_k} \underline{z}_{kj}.$$

Table 1
Observed Cholesterol Levels and Imputed Values

		Variable	1	2	3	4	5
Subset 2:	$n_2 = 7$		224	273	242	274	(231)
			231	252	267	299	(233)
			268	296	314	330	(303)
			284	288	268	261	(300)
			217	231	276	257	(238)
			209	200	269	233	(323)
			200	261	264	300	(279)
Subset 3:	$n_3 = 1$		193	189	(257)	232	211
Subset 4:	$n_4 = 12$		201	219	220	(231)	(172)
			202	186	253	(245)	(328)
			209	207	167	(208)	(194)
			212	253	225	(157)	(194)
			276	326	304	(300)	(376)
			163	179	199	(211)	(224)
			239	243	265	(238)	(246)
			204	203	198	(234)	(171)
			247	211	225	(224)	(215)
			195	250	272	(265)	(231)
			228	228	279	(276)	(259)
			290	264	260	(249)	(325)
Subset 5:	$n_5 = 1$		227	247	(215)	(267)	220
Subset 6:	$n_6 = 5$		250	269	(327)	(250)	(295)
			175	214	(250)	(210)	(210)
			260	268	(327)	(248)	(321)
			197	218	(235)	(251)	(258)
			248	262	(286)	(251)	(271)
Subset 7:	$n_7 = 2$		193	(209)	(219)	(230)	(255)
			256	(277)	(294)	(260)	(281)
Subset 8:	$n_8 = 1$		(284)	327	(287)	(336)	(309)

Note: Total sample size is $N = 65$.

Then

$$E(\underline{\bar{z}}_k) = B_k \underline{\mu},$$

$$\text{cov}(\underline{\bar{z}}_k) = n_k^{-1} (B_k \Sigma B_k'),$$

and the $\underline{\bar{z}}_k$ are independently distributed for $k = 1, \dots, K$. Applying the least squares theory, we minimize

$$\sum_{k=1}^K \text{tr } n_k (B_k \Sigma B_k')^{-1} [\underline{\bar{z}}_k - B_k \underline{\mu}] [\underline{\bar{z}}_k - B_k \underline{\mu}]'.$$

The solution for a given value of Σ is

$$\hat{\underline{\mu}} = \left[\sum_{k=1}^K n_k B_k' (B_k \Sigma B_k')^{-1} B_k \right]^{-1} \left[\sum_{k=1}^K n_k B_k' (B_k \Sigma B_k')^{-1} \underline{\bar{z}}_k \right]. \quad (2)$$

If a normal distribution is assumed, then the least squares estimator is also the maximum likelihood estimator. Little (1982) has suggested the use of the EM algorithm for this problem and claimed that the normal distribution assumption is not necessary. That is, estimators of $\underline{\mu}$ and Σ can be defined as the solution of the normal likelihood equations even if the underlying population is not normal. These estimators cannot then be considered maximum likelihood estimators, but only heuristic estimators that are consistent under certain general conditions. However, if a normal distribution is not assumed, then there is no justification in maximizing the normal likelihood equations to obtain estimators. An alternative heuristic estimator for Σ is given at the end of this section. The maximum likelihood estimator for Σ , assuming normality, are given from Srivastava (1985) as the solution of the following equation:

$$H = \sum_{k=1}^K n_k B_k' (B_k \Sigma B_k')^{-1} B_k - \sum_{k=1}^K B_k' (B_k \Sigma B_k')^{-1} V_k (B_k \Sigma B_k')^{-1} B_k = 0, \quad (3)$$

where

$$V_k = (\underline{z}_{k1} - B_k \underline{\mu}, \dots, \underline{z}_{k,n_k} - B_k \underline{\mu}) (\underline{z}_{k1} - B_k \underline{\mu}, \dots, \underline{z}_{k,n_k} - B_k \underline{\mu})'.$$

Methods for computing the solutions of (2) and (3) are given in Section 3.

Note: Alternate estimators for the covariance matrix can be defined heuristically without the normality assumption. For example $\hat{\Sigma}$ can be defined as the value of Σ that minimizes

$$\sum_{k=1}^K n_k^{-1} \text{tr} [(B_k \Sigma B_k')^{-1} V_k - n_k I_k]^2 \quad (4)$$

However, the covariance matrix must be positive definite; therefore any expression that is minimized must yield a positive definite solution. If one of the groups contains complete data, then (4) will be infinite for any singular matrix Σ ; hence, there will exist a minimum for (4) in the space of positive definite matrices. A similar argument holds for the maximum likelihood estimators.

2.3 Asymptotic Distribution of $\hat{\underline{\mu}}$.

From (2) it follows that $\hat{\underline{\mu}}$ is asymptotically normally distributed with mean $\underline{\mu}$ and covariance matrix

$$P = \left[\sum_{k=1}^K n_k B_k' (B_k \Sigma B_k')^{-1} B_k \right]^{-1}, \quad (5)$$

which can be estimated by \hat{P} obtained from P by substituting the $\hat{\Sigma}$ for Σ . Using this asymptotic theory, tests of significance and confidence regions (intervals) for $\underline{\mu}$ or linear combinations of $\underline{\mu}$ can be obtained. Alternatively, the likelihood ratio tests given by Srivastava (1985) may be used for testing the hypothesis $H: \underline{\mu} = \underline{0}$ against the alternative $A: \underline{\mu} \neq \underline{0}$. The likelihood ratio test rejects the null hypothesis H if

$$\lambda = \prod [|B_k \hat{\Sigma} B_k'| / |B_k \tilde{\Sigma} B_k'|]^{n_k/2} > \chi_{p, \alpha}^2,$$

where $\tilde{\Sigma}$ is the MLE of Σ under H and $\chi_{p, \alpha}^2$ is the upper 100 α % point of a chi-square distribution with p degrees of freedom.

2.4 Maximum Likelihood Estimates for Example 1

The maximum likelihood estimates for example 1 were obtained as:

$$\hat{\underline{\mu}} = \begin{pmatrix} 226.82 \\ 246.78 \\ 252.02 \\ 255.15 \\ 255.22 \end{pmatrix} \quad \text{and} \quad \hat{\Sigma} = \begin{pmatrix} 1809 & 1220 & 1033 & 873 & 913 \\ 1220 & 1642 & 992 & 1017 & 1121 \\ 1033 & 992 & 1438 & 718 & 1189 \\ 873 & 1017 & 718 & 1233 & 915 \\ 913 & 1121 & 1189 & 915 & 2508 \end{pmatrix}.$$

The estimated covariance matrix for the estimate of the mean vector is

$$P^{-1} = \begin{pmatrix} 28.05 & 18.78 & 15.96 & 13.46 & 14.08 \\ 18.78 & 25.67 & 15.42 & 15.84 & 17.51 \\ 15.96 & 15.42 & 24.19 & 11.24 & 19.31 \\ 13.46 & 15.84 & 11.24 & 23.33 & 15.38 \\ 14.08 & 17.51 & 19.31 & 15.38 & 54.77 \end{pmatrix}.$$

Inference on $\underline{\mu}$ can be made from the asymptotic distribution of the estimators given in Section 2.3.

2.5 Imputation

The imputation of the missing data can be made from the conditional distribution of the unobserved data given the observed data. That is define the matrices C_k for $k = 1, \dots, K$

to be the complements of B_k . That is for a $p_k \times p$ matrix B_k with ones as the (s, i_s) entries for $s = 1, \dots, p_k$ and 0's elsewhere, the matrix C_k is defined as the $(p - p_k) \times p$ matrix with ones in the (t, i_t) position and 0's elsewhere for $i_t \neq i_s$ for all $t = 1, \dots, (p - p_k)$ and $s = 1, \dots, p_k$. If the response vector y_{kj} corresponds to the j -th observation from subset k , then the actual observed response vector is $z_{kj} = B_k y_{kj}$ and the unobserved vector is $\hat{u}_{kj} = C_k y_{kj}$. The estimated value for the missing vector is given by

$$\hat{u}_{kj} = C_k \hat{\mu} + [C_k \hat{\Sigma} B_k'] [B_k \hat{\Sigma} B_k']^{-1} (z_{kj} - B_k \hat{\mu}) \tag{6}$$

Note that the estimated values for the missing vector have no random error. If the data is to be used at a subsequent analysis, with these imputed values, as if it were a complete data set, then the estimated error covariance matrix will be too small. The problem of underestimating the covariance matrix can be overcome by adding in an appropriate residual ϵ to the estimated value $\hat{\mu}_{kj}$. If the first subset of complete data is sufficiently large then the residual vectors for missing observations in subset k can be randomly drawn from the set of values

$$(C_k y_{1i} - C_k \hat{\mu}) - [C_k \hat{\Sigma} B_k'] [B_k \hat{\Sigma} B_k']^{-1} (B_k y_{1i} - B_k \hat{\mu}) \text{ for } i = 1, \dots, n_1. \tag{7}$$

Example 1 (continued):

The complete data set, including the imputed values based on (6) and (7) are given in Table 1 for subsets 2-8 with the imputed values in parenthesis.

3. COMPUTATIONAL PROCEDURES

Equations (2) and (3) can be solved iteratively. A procedure using a combined Newton-Raphson and steepest ascent method is given in Carter (1986) for a general case that includes linearly restricted means and covariances. The procedure is a generalization of the one given by Hartley and Hocking (1971). The method can be described as follows. For an initial choice of Σ , say Σ_0 , suppose

$$\Sigma = \Sigma_0 + \Lambda$$

is a solution. This expression is substituted into (3) and the equation is then expanded in a series involving only the linear terms of Λ . The following approximate solution for Λ results. Define

$$Q = \sum_{k=1}^K (D_k \otimes D_k - D_k \otimes F_k - F_k \otimes D_k),$$

where $A \otimes B$ denotes the kronecker product of two matrices A and B defined by $A \otimes B = (a_{ij} B)$,

$$D_k = B_k' (B_k \Sigma_0 B_k')^{-1} B_k,$$

and

$$F_k = B_k' (B_k \Sigma_0 B_k')^{-1} V_k (B_k \Sigma_0 B_k')^{-1} B_k.$$

For any matrix $A = (\underline{a}_1, \dots, \underline{a}_q)'$, we define $\text{vec}(A) = (\underline{a}_1', \dots, \underline{a}_q')'$. Then (3) can be written as approximately

$$Q \text{vec}(\Lambda) = \text{vec}(E),$$

where

$$E = \sum_{k=1}^K (D_k - F_k).$$

To insure the nonsingularity of Q , we shall write the solution for $\text{vec}(\Lambda)$ as

$$\text{vec}(\Lambda) = (Q + \lambda I)^{-1} \text{vec}(E), \quad (8)$$

where λ is allowed to vary with the algorithm but is initially set to a very small number. For a given value of Σ , $\hat{\mu}$ is obtained from (2) and then a value of Λ is obtained from (8) to produce an updated estimate for Σ . The procedure is then iterated until a desired level of convergence is reached.

The above method can be extended to more complex structured covariance matrices; however, the procedure does require the inversion of $Q + \lambda I$. For a large number of variables this matrix will be extremely large. In this instance the alternate method of solving (3) using the EM algorithm is preferable. Again the procedure is iterative, so calculations must be performed using the updated estimates of $\underline{\mu}$ and Σ at each iteration. For an initial choice of Σ say Σ_0 , define the complete predicted vector $\hat{y}_{kj} = B_k' \underline{z}_{kj} + C_k' \hat{\underline{\mu}}_{kj}$, where the predicted missing value $\hat{\underline{\mu}}_{kj}$ is given in (6). Then

$$\hat{\underline{\mu}} = (1/N) \sum_{k=1}^K \sum_{j=1}^n \hat{y}_{kj}$$

Define the matrix V by

$$V = \sum_{k=1}^K \sum_{j=1}^{n_k} (\hat{y}_{kj} - \hat{\underline{\mu}}) (\hat{y}_{kj} - \hat{\underline{\mu}})'$$

The updated estimate of Σ is then given by

$$\hat{\Sigma} = (1/N) [V + \sum_{k=1}^K n_k C_k' H_k C_k],$$

where H_k is the conditional variance of the incomplete data given the observed data for the k -th class defined by

$$H_k = C_k \Sigma C_k' - (C_k \Sigma B_k') (B_k \Sigma B_k')^{-1} (B_k \Sigma C_k').$$

The procedure is then iterated. The EM algorithm is advantageous for those situations where there exists simple closed form solutions for the likelihood equations in the complete data situations. If a Newton-Raphson procedure is necessary to solve the complete data likelihood equations then little is gained from the EM algorithm.

4. A REGRESSION MODEL

4.1 Incomplete Response Variables.

The model discussed in section 2 can be extended to handle the regression situation. The data is again partitioned into K subsets. Then the following regression model is formed:

$$Z_k = B_k' \beta A_k + \epsilon_k, \text{ for } k = 1, \dots, K,$$

where Z_k is a $p_k \times n_k$ matrix of observed values, β is a $p \times q$ matrix of unknown parameters, B_k is as defined in Section 2, A_k is the design matrix for the matrix Z_k and the columns of ϵ_k are independently distributed with mean 0 and covariance matrix $B_k \Sigma B_k'$. For a given Σ , the least squares estimator of β can be written from Carter (1986) explicitly as

$$\text{vec } \hat{\beta} = P^{-1} \text{vec}(E),$$

where

$$P = \sum_{k=1}^K n_k B_k' (B_k \Sigma B_k')^{-1} B_k \otimes A_k A_k', \tag{10}$$

$$E = \sum_{k=1}^K B_k' (B_k \Sigma B_k')^{-1} Z_k A_k'. \tag{11}$$

The maximum likelihood estimator of Σ is given by the same formula as (3), except that now

$$V_k = [Z_k - B_k \beta A_k][Z_k - B_k \beta A_k]'. \tag{12}$$

The asymptotic distribution of $\hat{\beta}$ can be written in the form

$$\text{vec}(\hat{\beta}) \sim N_{pq}(\text{vec}(\beta), P^{-1}). \tag{13}$$

Inference on the regression parameters can be made from this asymptotic distribution or from the likelihood ratio statistic given in Srivastava (1985).

4.2 Incomplete Explanatory Variables

In Section 3.1, the design matrices were assumed to be known completely. In some instances the explanatory variables can also be incomplete. If the explanatory variables are random, then these missing values can first be imputed for the explanatory variables given the observed data, using the procedure of Section 2 . Once imputed values for the explanatory variables are obtained then the method of Section 3.1 can be applied to estimate the regression parameters and to impute the missing response variables.

4.3 The Likelihood Ratio Test.

The likelihood ratio procedure can be used to determine if the variables in the model are significant. To test the hypothesis

$$H: \beta = \beta_1 F \text{ vs } A: \beta \neq \beta_1 F,$$

for F an $m \times q$ matrix of full rank, the estimates of Σ are obtained under the null hypothesis ($\tilde{\Sigma}$) and under the alternate hypothesis ($\hat{\Sigma}$). The null hypothesis is rejected at the α level of significance if

$$-2 \ln \lambda > \chi^2_{(q-m)p; \alpha},$$

where

$$\lambda = \prod_{k=1}^K |B_k \hat{\Sigma} B_k'|^{n_k/2} / |B_k \tilde{\Sigma} B_k'|^{n_k/2}. \quad (14)$$

5. ESTIMATING A CHANGE POINT

Consider a sequence of observations $y_j, j = 1, \dots, N$, with expected values $E(y_j) = \underline{\mu}_j$. Srivastava and Worsley (1986) have given a procedure for estimating the point of change of the mean vectors $\underline{\mu}_j$. It is first assumed that the change occurs at some point r . Then the following hypothesis is tested.

$$H: \underline{\mu}_1 = \dots = \underline{\mu}_N$$

$$A: \underline{\mu}_1 = \dots = \underline{\mu}_r \neq \underline{\mu}_{r+1} = \dots = \underline{\mu}_N.$$

The likelihood ratio statistic is then calculated as λ_r , for $r = 1, \dots, N - 1$. The estimated point of change is that value of r that yields the maximum value of λ_r .

The existence of incomplete data poses no problems for estimating the change point. The linear model is set up as for the complete data case, then the observations are grouped into the K subsets. Suppose that the observed portion of y_j is z_{ki} . Then under the alternate hypothesis for a given r , $\hat{\Sigma}$ the estimate for Σ is given from (3) for the regression model defined in (9)-(12), where the parameter matrix β is defined as

$$\beta = (\underline{\mu}_1, \underline{\mu}_2)$$

and the design matrix for the k -th subset is defined by

$$A_k = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \end{bmatrix},$$

where the i -th column of A_k has a one in the first row if observation z_{ki} corresponds to the vector y_j and $j \leq r$ and zero otherwise. Under the null hypothesis the population mean vector is considered the same for all N observations; hence, Σ the estimate for Σ is given from (2) and (3) for the one population mean problem. The likelihood ratio statistic is obtained from (14).

Modifications of this procedure are possible. For example the vectors y_j for $j = 1, \dots, N$ could be sample means for N sampling time points. Multiple change points can be obtained by repeating the procedure on each section of the data. For 50 observations, if the change point occurs at point 20 then the procedure is repeated for points 1-20 and 21-50.

6. STRUCTURED COVARIANCE MATRICES

For longitudinal studies the error vectors over time may not be arbitrary, but may follow a time series model. If such a model can be assumed, then the number of parameters to be estimated is reduced. A stationary time series would assume that the covariance matrix Σ can be written as

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \dots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 \dots & \rho_{p-2} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \rho_{p-1} & \rho_{p-2} & \dots \dots \rho_1 & 1 \end{pmatrix} \cdot \tag{15}$$

Further models can be obtained. The correlations ρ_j can be structured. For example ρ_j can be set equal to $\rho^{|j|}$. The likelihood equations can be solved using the Newton-Raphson technique. Carter (1986) considered the case where the covariance matrix can be written as $\text{vec}(\Sigma) = G\underline{\gamma}$ for some matrix G . By defining $\gamma_i = \sigma^2\rho_i$ for $i = 1, \dots, p - 1$ and $\gamma_p = \sigma^2$, then the covariance matrix for the stationary time series can be expressed in this linearly restricted form. For example for $p = 3$ we have

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{21} \\ \sigma_{22} \\ \sigma_{23} \\ \sigma_{31} \\ \sigma_{32} \\ \sigma_{33} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}$$

The estimate of Σ can be solved numerically from the likelihood equation $G'H = 0$, where H is defined in (3). Numerically the Newton-Raphson algorithm from Section 3 can be employed with the modification that the estimate for $\underline{\gamma}$ at each iteration is given by

$$\hat{\underline{\gamma}} = (G'QG + \lambda I)^{-1}G' \text{vec}(E).$$

REFERENCES

- CARTER, E.M. (1986). The analysis of a generalized multivariate linear model. Technical Report, University of Guelph.
- DAVID, M., LITTLE, R.J., SAMUHEL, M.E., and TRIEST, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.
- DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 39, 1-38.
- DRAPER and SMITH (1981). *Applied Regression Analysis*. New York: Wiley.
- GREENLEES, W.S., REECE, J.S., and ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- HARTLEY, H.O., and HOCKING, R.R. (1971). The analysis of incomplete data (with discussion). *Biometrics*, 27, 783-823.
- HECKMAN, J.D. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimation for such models. *Annals of Economic and Social Measurements*, 5, 475-492.
- LITTLE, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- RUBIN, D.B., and SZATROWSKI, T.H. (1982). Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm. *Biometrika*, 69, 657-660.
- SRIVASTAVA, M.S. (1985). Multivariate data with missing observations. *Communications in Statistics Theory and Methods*, 14, 775-792.
- SRIVASTAVA, M.S., and WORSLEY, K.J. (1986). Likelihood ratio tests for a change in the multivariate mean. *Journal of the American Statistical Association*, 81, 199-204.
- WEI, L.J., and LACHIN, J.M. (1984). Two sample asymptotically distribution free tests for incomplete multivariate observations. *Journal of the American Statistical Association*, 79, 653-661.
- WOOLSON, R.F., and LEEPER, J.D. (1980). Growth curve analysis of complete and longitudinal data. *Communications in Statistics, Theory and Methods*, 9, 1491-1513.
- WOOLSON, R.F., LEEPER, J.D., and CLARKE, W.R. (1978). Analysis of incomplete data from longitudinal and mixed longitudinal studies. *Journal of the Royal Statistical Society, Ser. A*, 141, 242-252.

Statistical Editing and Imputation for Periodic Business Surveys

M.A. HIDIROGLOU and J.-M. BERTHELOT¹

ABSTRACT

For periodic business surveys which are conducted on a monthly, quarterly or annual basis, the data for responding units must be edited and the data for non-responding units must be imputed. This paper reports on methods which can be used for editing and imputing data. The editing is comprised of consistency and statistical edits. The imputation is done for both total non-response and partial non-response.

KEY WORDS: Periodic survey; Statistical editing; Total/partial non-response; Imputation.

1. INTRODUCTION

Data are routinely collected by large organizations such as Statistics Canada based on properly designed sample surveys. If such data are collected on a periodic basis from the same sampling unit, there are several possibilities which will occur with respect to the data consistency (quality) over a given time period. The sampling unit may report the data faithfully with no dramatic departure in continuity ("smoothness") as time progresses. The data may be reported faithfully, with questionable jumps between two time periods. The sampling unit may not report all the requested data items: this is known as partial non-response. The sampling unit may report data sporadically with breaks of total non-response for some periods. These can occur simultaneously in a periodic survey which collects required data from a large number of sampling units.

The problems which will be addressed in this article are the editing and imputation of data for sampling units that are contacted on a periodic basis by a surveying organization. The methods discussed are general for data of a multivariate nature composed of both quantitative and qualitative variables. The editing will include consistency and statistical edits.

For quantitative data, consistency edits ensure that linear combination of the data fields within a given time period satisfy given requirements. For qualitative data, consistency edits ensure that variables correspond to well defined values.

Statistical edits are used to isolate sampling units which may report some of their quantitative data fields in an inconsistent manner either from time period to time period or within a specific time period. Units with unusually high or low values will be termed "outliers". The identification of "outliers" is extremely important in an ongoing survey for two reasons. First, they influence statistics of the data set which may be for instance totals. This point has been studied by Hidiroglou and Srinath (1981). Second, the imputation of quantitative data for non-response units for periodic business surveys is usually based on trends or means: the removal of outlier units from the computation of these trends or means, will produce statistics that are not contaminated with these observations. For units which have partial non-response, data must be imputed for the missing fields.

For large data sets, where timely release of the summary information is crucial, the editing and the imputation of data should be automatic and computer handled given some well specified rules. This is in agreement with Gentleman and Wilk (1975), and Fellegi and Holt (1976).

¹ M.A. Hidiroglou and J.-M. Berthelot, Business Survey Methods Division, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

2. EDITING PERIODIC DATA

2.0 Consistency Edits

For a given unit i and time period t , let $\underline{x}_i(t)$ represent the vector of data which is to be collected. The vector $\underline{x}_i(t)$ may be decomposed into a series of elementary vectors for which independent editing and imputation are required.

That is,
$$\underline{x}_i(t) = (\underline{x}_i^{(1)}(t), \dots, \underline{x}_i^{(P)}(t))$$

where
$$\underline{x}_i^{(p)}(t) = (x_{i1}^{(p)}(t), \dots, x_{ik_p}^{(p)}(t))$$

for $i=1, \dots, n; p=1, \dots, P; t=1, \dots, T$

and k_p is the number of variables in the p :th elementary vector.

For each elementary vector $\underline{x}_i^{(p)}(t)$, the consistency edits may be represented as

$$A^{(p)}(\underline{x}_i^{(p)}(t))' \leq (\underline{c}^{(p)})'$$

where $A^{(p)}$ is a ℓ_p by k_p matrix representing the rules that the elements of the elementary vector $\underline{x}_i^{(p)}(t)$ must obey, and $\underline{c}^{(p)}$ is a 1 by ℓ_p vector which represents the constraints. This formulation allows one to define consistency edits for both qualitative and quantitative variables. For qualitative variables, the consistency edits could be used to check if the variables correspond to well-defined values. For quantitative variables, the consistency edits can check if certain variables are not larger (or smaller) than other variables or that a linear combination is equal to (or greater than or less than) a given variable.

2.1 Statistical Edits

Given that data are reported periodically, the problem is to isolate outlying observations within the time series. In the present context, an outlying observation i , will be defined as one whose trend for the current period to a previous period, for given variables of the element vector $\underline{x}_i(t)$, differs significantly from the corresponding overall trend of other observations belonging to the same subset of the population. Statistical edits can also be applied within a time period, by comparing the ratios of two correlated variables amongst themselves, within a given subset of the population. In this article, the statistical edit will only be discussed in terms of the trend between time periods. Similar, somewhat imprecise but working definitions of outliers have also been given by other authors, for example:

GRUBBS (1969) says that "An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs."

GUMBEL (1960) says: "The outliers are values which seem either too large or too small as compared to the rest of the observations."

KENDALL and BUCKLAND (1957, p. 209), write: "In a sample of n observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are from a different population, or that the sampling technique is at fault. Such values are called outliers. Tests are available to ascertain whether they can be accepted as homogeneous with the rest of the sample."

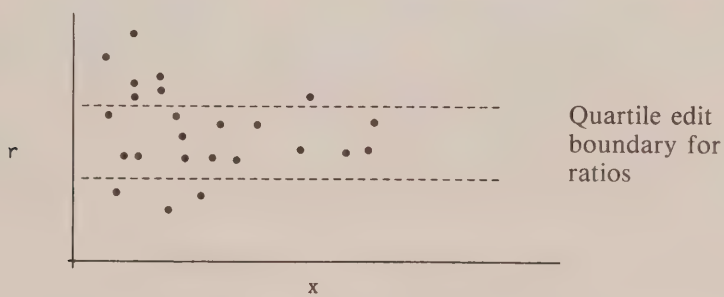
2.1.1 Review of Some Methods Currently Used

Methods for detecting outliers have been proposed by Dixon (1953), Grubbs (1969), Tietgen and Moore (1972), and Prescott (1978) to mention a few. Most of the test procedures for outlier detection proposed by these authors consider the problem as one of hypothesis testing. In the simplest cases, the null hypothesis is that the sample comes from a normal distribution with unspecified mean and variance, while the alternative hypothesis is that one or more of the observations come from a different distribution. Percentage points of a test statistic may be determined under the null hypothesis and compared with computed values of the test statistic in particular applications. Applying these methods to periodic data from large surveys presents problems for the following reasons. First, the assumption of normality of trends from one period to another may not hold. Second, these traditional methods require the existence of tables for determining critical values which define rejection regions. The method which we will propose in Section 2.1.2 does not have the above mentioned disadvantages. It can be easily implemented on the computer, does not require the assumption of normality, and does not make use of tables.

In our specific context, and given elements of the vectors $\underline{x}_i(t)$ and $\underline{x}_i(t + 1)$, denote as $x_i(t)$ and $x_i(t + 1)$ the responses for two consecutive periods for a given unit, where $i=1, \dots, n$. Denote as r_i the ratio of current period data to previous period data. One method which is known as the range edit, is to simply define fixed upper and lower bounds based on experience for comparison purposes. Ratios found outside these bounds are declared as outliers. A major drawback with this method is that the definition of outlier is too subjective and does not make use of the distribution of the ratios.

A method that attempts to make use of the distribution of the ratios is the Chebychev inequality edit. This edit is constructed by computing the lower bound as $\bar{r} - ks_r$ and the upper bound as $\bar{r} + ks_r$ where $\bar{r} = \sum_{i=1}^n r_i/n$ and $s_r^2 = \sum_{i=1}^n (r_i - \bar{r})^2/(n - 1)$. This edit has two main drawbacks. First, the choice of k is subjective and can result in having an edit that cannot detect any outliers. This last point has been demonstrated by Wilkinson (1982). Second, "large" outliers may hide "smaller" outliers. This effect is known as the masking effect.

An improvement to this method has been the use of quartiles and interquartile distances rather than the use of mean and standard error to come up with the upper and lower bounds. In this case, the edit is constructed by computing the lower bound as $r_M - k D_{r_{Q1}}$ and the upper bound as $r_M + k D_{r_{Q3}}$ where r_M is the median of the ratios, $D_{r_{Q1}}$ is the distance between the first quartile and the median, and $D_{r_{Q3}}$ is the distance between the third quartile and the median. Since the quartiles are not affected by the tails of the distribution, it greatly alleviates the masking effect problem. However, this method has two drawbacks. First, in some very specific circumstances, it is possible that the outliers on the left tail of the distribution are undetectable. Second this method does not take into account the fact that in most of the periodic business surveys, the variability of ratios for small businesses is larger than the variability of ratios for large businesses (Sugavanam 1983). This fact is expressed by the following graph:



This drawback has the effect of identifying too many small units as outliers and not enough large units. This effect will be referred to as the "size masking effect".

2.1.2 Proposed Procedure

For two occasions t and $t + 1$, the overall trend for the data pair given by

$$(x_i(t), x_i(t + 1)), i = 1, \dots, n$$

is

$$R = \sum_{i=1}^n x_i(t + 1) / \sum_{i=1}^n x_i(t).$$

Now, R may be expressed as

$$R = \sum_{i=1}^n I_i r_i$$

where

$$I_i = x_i(t) / \sum_{i=1}^n x_i(t)$$

and

$$r_i = x_i(t + 1) / x_i(t).$$

I_i is a measure of the relative importance of the i :th unit amongst the n units at time t . The individual trends r_i must be transformed in order to ensure that outliers are detected at both tails of the distribution. This transformation is:

$$s_i = \begin{cases} 1 - r_M/r_b & \text{if } 0 < r_i < r_M \\ r_i/r_M - 1 & \text{if } r_i \geq r_M \end{cases}$$

where r_M is the median of the ratios.

In order to bring in the magnitude of the data, the following transformation is required (Berthelot 1983):

$$E_i = s_i \{ \text{Max } (x_i(t), x_i(t + 1)) \}^U$$

where $0 \leq U \leq 1$. The E_i 's will be referred to as effects and the exponent U in the transformation provides a control on the importance associated with the magnitude of the data. This transformation allows us to place more importance on a small change associated with a "large" unit as opposed to a large change associated with a "small" unit. The values of the median and quartiles as used by Sande (1981) will be applied to the transformed, E_i 's, in order to detect potential outliers. Denoting as E_{Q1} , E_M and E_{Q3} as the first quartile, the median and the third quartile respectively, define the following two deviations:

$$d_{Q1} = \text{Max } (E_M - E_{Q1}, |AE_M|),$$

$$d_{Q3} = \text{Max } (E_{Q3} - E_M, |AE_M|).$$

Outliers will be defined as all those units whose associated effect E_i lies outside the interval $(E_M - Cd_{Q1}, E_M + Cd_{Q3})$. The purpose of the AE_M term is to avoid difficulties which arise when $E_M - E_{Q1}$ or $E_{Q3} - E_M$ are very small. That is, the problem which may arise when the effects E_i are clustered around a single value with one or two modest deviations may produce false outliers. The parameter C controls the width of the acceptance interval. The parameter U controls the shape of the curve defining upper and lower boundaries. The effect of increasing U is to attach more importance with fluctuations associated with the larger observations. A value of 0.05 is suggested for A as it has proved to be adequate in practice.

2.1.3 Treatment For Outliers

Once units have been identified as possible outliers, they are flagged as such and brought to the attention of the survey takers. A decision must then be taken on how these abnormal observations are treated. Their existence may have arisen as a result of several factors. These factors include measurement error, incorrect interpretation of the questionnaire by the responding unit, or intrinsic variability of the population being surveyed. For units which have measurement error due to incorrect transcription of the data or incorrect responses, a simple follow-up will clear up the majority of these errors. For units which display intrinsic variability as a result of rapid growth, the reported values are correct but dominate too much the resulting summary tables. For those units, techniques, which reduce the sampling weight as suggested by Hidirolou and Srinath (1981) or change the values themselves as suggested by Ernst (1980), must be used in order to accomodate (minimize) the effect of outlying observations. For units having unrepresentative data which cannot be verified, their data must be substituted with other data based on imputation techniques. The different kinds of corrective actions taken on outlying units must be flagged as well.

3. IMPUTING PERIODIC DATA

The information collected by periodic business surveys, such as sales and employment are collected via samples using mail questionnaires or telephone interviews. Non-responding units are followed up as much as possible within allotted budgets in order to improve the response rates. The follow-up is usually done by mail in the case of the smaller to medium sizes non-responding companies and by telephone for the larger or dominating companies. Although following up delinquent companies improves response rates for a given reference period, there will be nevertheless, a group of non-responding companies which may be classified into either hard-core or late respondents. Hard-core non-respondents are units which require a great deal of persuasion to respond, if at all. Late respondents are units which respond late with respect to the survey's reference period either because they do not mail back their questionnaire on time or because they need to be prompted by a follow-up questionnaire. The non-responding units must therefore be imputed in order to make up for their contribution to the particular estimator being used by the survey. In the case of Monthly Business Surveys, such as the Monthly Retail Trade Survey, totals (e.g., sales) are being estimated. Imputation procedures can also be used to generate values for units declared as outliers. These imputed values can be used in lieu of these outlying observations, if no valid explanation can be provided for their presence.

The units with no response whatsoever, will be termed as total non-respondents and those with some, but not all, required data items, will be termed partial non-respondents. Desirable features of an imputation system should include the following properties (Berthelot and Hidirolou 1982):

- it must automatically determine the most reasonable imputation procedure possible under the existing circumstances,
- the imputation cell, the level at which the computation of trends and means (medians) is performed, will usually correspond to the finest level of stratification of the sample,
- a minimum number of units must participate in the computation of trends or means (medians), otherwise, the imputation cells are automatically collapsed (using a pre-determined pattern), until the minimum requirement has been satisfied,
- it will recognize through the use of status codes that there are units which must not be imputed. These include seasonal units during the period that they are not operating, units temporarily out of business, or units which are no longer active,
- births which have no previous business history will have their data imputed using the means (medians) of similar responding births,
- units will be re-imputed for a number of periods previous to the current period: this is done in order to improve the strength of the imputations if the previous periods have been updated with data,
- backward imputations will be applied to units which have been continuously imputed using a forward imputation procedure as soon as a good response is obtained for a given period,
- imputation status codes will be associated with imputed units in order to provide a history of the procedure used for imputation,
- the ranking for imputing non-responding units is as follows: trends (monthly, quarterly, annual), means (medians) with the most recent trends being given priority. For instance, in the case of a monthly system, monthly trends are used for units which have data (response or imputed) in the month prior to the one to be imputed. Annual trends are used mostly for units which are seasonal and which fail to provide a response as they emerge from their out of season period and for which a last year value existed for the month to be imputed. Imputations based on the trends are obtained by multiplying the trends by the unit's last month or last year value. In the event that trends cannot be applied, the mean (median) of the cell is used as an imputation.

In order to formalize the preceding paragraphs in a mathematical fashion, let the number of units which are expected to respond for a given cell and given month be n . Let the number of non-respondents with total non-response be n_3 , the number of respondents with total response be n_1 and the number of respondents with partial response be n_2 . It is assumed that the sample design is stratified with the sampling being simple random without replacement. Let the size for the follow-up sample of the non-respondents be m_3 ($2 \leq m_3 \leq n_3$, with m_3 having been selected from n_3 according to a randomized mechanism). Note that $n_4 = n - \sum_{i=1}^3 n_i$ units are not expected to provide any response to the survey process for a number of possible reasons. At a time t , they may be out of season, inactive, dead, or out of scope to the survey. For these units, the system will automatically associate zero values for all relevant fields in the given period.

The imputation process will then be done in several different ways according to the type of non-response.

3.0 Total Non-Response

The imputation process for the total non-respondents will first be discussed. Bearing in mind that either the whole vector $x_i(t)$ or that some of its elementary vectors as given in

Section 2.0 must be totally imputed, denote as $(x_{i1}(t), \dots, x_{ip}(t))$ one of the elementary vector within $x_i(t)$ where the editing and imputation process is independent from other elementary vectors within $x_i(t)$. Assuming that

$$x_{ip}(t) \geq \sum_{j=1}^{p-1} x_{ij}(t),$$

(which implies that the sum of the first $p-1$ data elements of the elementary vectors are smaller than the p :th datum element, the total) $x_{ip}(t)$ will first be imputed as

$$I_{ip}^{(1)}(t) = \sum_{k=1}^6 [z_{ip}^{(k)}(t) \delta_i^{(k)}]$$

where $\delta_i^{(k)}$ refers to the procedure used for imputation and $z_{ip}^{(k)}$ is the associated imputed value. One of the six $\delta_i^{(k)}$ values will be one and the other five must be zero ($\sum_{k=1}^6 \delta_i^{(k)} = 1$). The imputed $z_{ip}^{(k)}(t)$ values will be as follows:

$$z_{ip}^{(1)}(t) = [\sum_{r \in s_1} w_r x_{rp}(t) / \sum_{r \in s_1} w_r x_{rp}(t-1)] x_{ip}(t-1),$$

$$z_{ip}^{(2)}(t) = [\sum_{r \in s_2} w_r x_{rp}(t) / \sum_{r \in s_2} w_r x_{rp}(t-Q)] x_{ip}(t-Q),$$

$$z_{ip}^{(3)}(t) = [\sum_{r \in s_3} w_r x_{rp}(t) / \sum_{r \in s_3} w_r x_{rp}(t-1)] x_{ip}(t-1),$$

$$z_{ip}^{(4)}(t) = [\sum_{r \in s_4} w_r x_{rp}(t) / \sum_{r \in s_4} w_r x_{rp}(t-Q)] x_{ip}(t-Q),$$

$$z_{ip}^{(5)}(t) = [\sum_{r \in s_5} w_r x_{rp}(t) / \sum_{r \in s_5} w_r],$$

$$z_{ip}^{(6)}(t) = [\sum_{r \in s_6} w_r x_{rp}(t) / \sum_{r \in s_6} w_r],$$

w_r = inverse selection probability of unit r for the given cell. The subsets s_i ($i=1, \dots, 6$), will be determined by selecting the units which have provided a response for the p :th variable at time t and which have passed the edits. The conditions for each subset is

s_1 = all units which have provided edited responses between times t and $t-1$,

s_2 = all units which have provided edited responses between times t and $t-Q$,

s_3 = units in the follow-up subsample which have provided edited responses between times t and $t-1$,

s_4 = units in the follow-up subsample which have provided edited responses between times t and $t-Q$,

s_5 = all units which have provided edited responses at time t ,

s_6 = units in the follow-up subsample which have provided edited responses at time t .

The choice of the imputation procedure will be governed by the following considerations.

- (i) Procedures 1 (or 2) will be used if there is a response or imputed value at time $t-1$ (or $t-Q$) and that it is believed that the trends for the non-respondents is the same as the one for the respondents, within the given cell,
- (ii) Procedures 3 (or 4) will be used if there is a response or imputed value at time $t-1$ (or $t-Q$) and that it is believed that the trends for the non-respondents differs from the one for the respondents within the given cell.
- (iii) Procedure 5 will be used if there is no response at either times $t-1$ or $t-Q$ and that it is believed that the mean of the non-respondents is equal to the mean of the respondents within the given cell,
- (iv) Finally, procedure 6 will be used if there is no response at either times $t-1$ or $t-Q$ and that it is believed that the means of the respondents and non-respondents are different.

The choices between the different procedures can be made using decision tables which determine the conditions and, given the condition, choose the best imputation procedure according to pre-determined rules. Once that $x_{ip}(t)$ has been imputed for an elementary vector, its remaining components can be imputed using the procedures for partial non-response.

3.1 Partial Non-Response

For an elementary vector $(x_{i1}(t), x_{i2}(t), \dots, x_{ip}(t))$ which is part of $\underline{x}_i(t)$, let δ_{ij} be the indicator variable which is equal to 1 if $x_{ij}(t)$ is present and zero otherwise at time t . Some additional notation is introduced at this point in order to ease the development. To this end, define

$$s_{i,R}(t-1) = \sum_{j=1}^{p-1} \delta_{ij} x_{ij}(t-1)$$

= the sum of responses at time $t-1$, for which
there is a response at time t

$$s_{i,NR}(t-1) = \sum_{j=1}^{p-1} (1-\delta_{ij}) x_{ij}(t-1)$$

= the sum of responses at time $t-1$, for which
there is no response at time t ,

$$s_{i,R}(t) = \sum_{j=1}^{p-1} \delta_{ij} x_{ij}(t).$$

The partial imputation will be based on the assumptions that $x_{ip}(t) \geq \sum_{j=1}^{p-1} x_{ij}(t)$ and that the distribution of the elements within $\underline{x}_i(t)$ is similar to the distribution of the elements within $\underline{x}_i(t-1)$. Two separate cases will be discussed.

Case 1: Parts of the elementary vector missing and $x_{ip}(t)$ present

Two subcases are possible: $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$ or $x_{ip}(t) > \sum_{j=1}^{p-1} x_{ij}(t)$.

$$(i) \quad x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$$

If all the elements of $x_i(t)$ excluding $x_{ip}(t)$ are missing, that is $\sum_{j=1}^{p-1} \delta_{ij} = 0$, then we must have that $s_{i, NR}(t) = x_{ip}(t)$. If some of the elements of $x_i(t)$ excluding $x_{ip}(t)$ are missing, that is $\sum_{j=1}^{p-1} \delta_{ij} > 0$, then $s_{i, NR}(t) = x_{ip}(t) - s_{i, R}(t)$.

$$(ii) \quad x_{ip}(t) > \sum_{j=1}^{p-1} x_{ij}(t)$$

If all the elements of $x_i(t)$ excluding $x_{ip}(t)$ are missing, then $s_{i, NR}(t) = s_{i, NR}(t-1) x_{ip}(t) / x_{ip}(t-1)$. If some of the elements of $x_i(t)$ excluding $x_{ip}(t)$ are missing, the choice of $s_{i, NR}(t)$ is not so obvious. In any event, one must have that $s_{i, R}(t) + s_{i, NR}(t) < x_{ip}(t)$. To this end, four separate possible imputations for $s_{i, NR}(t)$ will be given in order of preference.

(a) $s_{i, NR}(t) = [s_{i, NR}(t-1) + s_{i, R}(t-1)] x_{ip}(t) / x_{ip}(t-1) - s_{i, R}(t)$ provided that $s_{i, NR}(t) \geq 0$. Note that the condition $x_{ip} > \sum_{j=1}^{p-1} x_{ij}(t)$ is met if $s_{i, NR}(t) \geq 0$.

(b) $s_{i, NR}(t) = s_{i, NR}(t-1) [s_{i, R}(t) / s_{i, R}(t-1)]$

(c) $s_{i, NR}(t) = s_{i, NR}(t-1) [x_{ip}(t) / x_{ip}(t-1)]$

(d) $s_{i, NR}(t) = x_{ip}(t) - s_{i, R}(t)$.

The preferred imputation will be the first one that does not violate the inequality condition. For all the above cases, the imputed (actual values) will then be

$$I_{ij}^{(2)}(t) = (1 - \delta_{ij}) [s_{i, NR}(t) / s_{i, NR}(t-1)] x_{ij}(t-1) \\ + \delta_{ij} x_{ij}(t); j=1, \dots, p-1$$

Case 2: Parts of the elementary vector missing and $x_{ip}(t)$ is missing

As in case 1, two subcases are possible:

$$(i) \quad x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$$

If $\sum_{j=1}^{p-1} \delta_{ij} = 0$, then $s_{i, NR}(t) = I_{ip}^{(1)}(t)$ where $I_{ip}^{(1)}(t)$ has been obtained using the imputation for total non-response. The imputation $I_{ij}^{(2)}(t)$ is then used. If $\sum_{j=1}^{p-1} \delta_{ij} > 0$, $I_{ij}^{(2)}(t)$ will be used provided that $s_{i, NR}(t) = I_{ip}^{(1)}(t) - s_{i, R}(t) \geq 0$. Otherwise, the following imputation must be used

$$I_{ij}^{(3)}(t) = (1 - \delta_{ij}) [s_{i, NR}(t) / s_{i, NR}(t-1)] x_{ij}(t-1) \\ + \delta_{ij} x_{ij}(t); j=1, \dots, p-1$$

and $I_{ip}^{(1)}(t)$ is replaced by $I_{ip}^{(3)}(t) = \sum_{j=1}^{p-1} I_{ip}^{(3)}(t)$

(ii) $x_{ip}(t) > \sum_{j=1}^{p-1} x_{ij}(t)$

For this case, the $x_{ip}(t)$ in case 1(ii) is replaced by $I_{ip}^{(1)}(t)$ and the methods given for this case are used, provided that the above inequality condition is satisfied. If the condition cannot be met, $I_{ip}^{(3)}(t)$ must be used and $I_{ip}^{(1)}(t)$ is replaced by $I_{ip}^{(3)}(t) = \sum_{j=1}^{p-1} I_{ip}^{(3)}(t)$.

If the assumption, that the distributions of the data elements of vectors $x_i(t)$ and $x_i(t-1)$ is similar, does not hold, then each individual element must be imputed using procedures for imputation for total non-response. These imputations must then be adjusted in order to satisfy the inequality requirement $x_{ip} \geq \sum_{j=1}^{p-1} x_{ij}$. Hence, for example, for case 1(i), we would have for $\sum_{j=1}^{p-1} \delta_{ij} = 0$,

$$I_{ij}^{(4)}(t) = [x_{ip}(t) / \sum_{j=1}^{p-1} I_{ij}^{(1)}(t)] I_{ij}^{(1)}(t)$$

and for $\sum_{j=1}^{p-1} \delta_{ij} > 0$

$$I_{ij}^{(4)}(t) = (1 - \delta_{ij}) \left[\frac{x_{ip}(t) - \sum_{j=1}^{p-1} \delta_{ij} x_{ij}(t)}{\sum_{j=1}^{p-1} (1 - \delta_{ij}) I_{ij}^{(1)}(t)} \right] + \delta_{ij} x_{ij}(t); j = 1, \dots, p-1.$$

Similarly, cases 1(ii) and 2, could be developed using the imputed values $I_{ij}^{(1)}(t)$.

4. CONCLUSION

For periodic business surveys, it is important to have computer systems which can quickly and accurately monitor the flow of in-coming data in terms of its quality. Conversely, for expected data that are not coming in, the system should impute as well as possible for the non-response given some well specified rules.

The editing will cause the flagging of records in possible error. These errors can be termed as critical and non-critical. All errors should be corrected by either reviewing the questionnaires or checking their authenticity with the respondent. If this is not possible on account of time or budgetary constraints, the most critical errors must be corrected. Given that the errors have been taken care of, the next step of the processing is to impute for the non-respondents. Diagnostic summaries of the actions (edits or imputations) taken by the system, should be printed out in order to inform the survey analyst on the status of his data.

REFERENCES

- BERTHELOT, J.-M., and HIDIROGLOU, M.A. (1982). Specifications for imputations in the retail trade survey. Technical report, Statistics Canada.
- BERTHELOT, J.-M. (1983). Wholesale-retail redesign, statistical edit proposal. Technical Report, Statistics Canada.
- DIXON, W.G. (1953). Processing data for outliers. *Biometrics*, 9, 74-89.

- ERNST, L.R. (1980). Comparison of estimators of the mean which adjust for large observations. *Sankhya*, 42, 1-16.
- FELLEGI, I.P., and HOLT, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.
- GENTLEMAN, J.F., and WILK, M.B. (1975). Detecting outliers, II. Supplementing the direct analysis of residuals. *Biometrics*, 31, 387-410.
- GRUBBS, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11, 1-21.
- GUMBEL, E.J. (1960). Discussion on "Rejection of outliers" by Anscombe, F.J. *Technometrics*, 2, 165-166.
- HIDIROGLOU, M.A., and SRINATH, K.P. (1981). Some estimators of population totals form a simple random sample containing large units. *Journal of the American Statistical Association*, 76, 690-695.
- KENDALL, M.G., and BUCKLAND, W.R. (1957). *A Dictionary of Statistical Terms*. New York: Hafner.
- PRESCOTT, P. (1978). Examination of the behaviour of tests for outliers when more than one outlier is present. *Applied Statistics*, 27, 10-25.
- SUGAVANAM, R. (1983). A statistical edit for change. Technical Report, Statistics Canada.
- SANDE, I.G. (1981). Estimation in the revised ISPI. Technical Report, Statistics Canada.
- TIETGEN, G.L., and MOORE, R.H. (1972). Some Grubbs - type statistics for the detection of several outliers. *Technometrics*, 55, 583-598.
- WILKINSON, R.G. (1982). An outlier identification technique designed for the Business Finance Annual Survey. Technical Report, Statistics Canada.

Practical Criteria for Definition of Weighting Classes

VICTOR TREMBLAY¹

ABSTRACT

When the technique of adjustment using weighting classes is applied to compensate for the effect of non-response, several questions arise that call for precise and quantified answers: How does the choice of the variables used for definition of the classes affect total root-mean-square error, in particular non-response bias and sampling variance? What rule and what procedure should be followed in choosing the adjustment variables? On the basis of what criterion can the optimal sizes for the weighting classes be established? Finally, when this procedure is applied to compensate for non-response with respect to specific elements of a questionnaire, how can strongly correlated ancillary variables be used effectively when they themselves are affected by non-response? This article is addressed to those professionals working at a practical level who are seeking guidelines.

KEY WORDS: Adjustment for non-response; Weighting classes; Poststratification; Non-response bias.

1. INTRODUCTION

The problem of adjustment for non-response through creation of weighting classes is clearly related to that of determination of poststratification criteria. Kish (1978) stated that there was an urgent need for research in this area, noting that, in terms of advantages and disadvantages, the final effect of this type of weighting is often unknown. At the same time, Platek, Singh and Tremblay (1978) developed mathematical expressions for the bias and the variance of the estimators resulting from adjustment using weighting classes. Their model, which was based on the response-probability concept, was developed further recently by Platek and Gray (1983). During the same period, Bailer, Bailey and Corby (1978) described the theoretical and empirical research undertaken at the US Bureau of the Census. They end their presentation by emphasizing the importance and the necessity of developing solid theoretical foundations for the methods of adjustment for non-response. More recently, the Panel on Incomplete Data (1983) provided a particularly concise and complete description of the practical implications of adjustment through weighting and stressed the conclusions reached by Oh and Scheuren (1983) following a simulation study. Chapman (1983) analysed a number of procedures that could be used to identify the most relevant variables for effective construction of weighting classes.

This article continues along the same lines as these research efforts by attempting to define some rules for application of this adjustment procedure starting from theoretical foundations. The single example used for illustration throughout this text is very specific, but the reader will no doubt be able to identify much more varied and rich application possibilities.

2. ILLUSTRATION OF THE TECHNIQUE

Let us take as our example the measurement of voters' intentions, a very real and frequently encountered problem. All of the data used in this text comes from the fall 1985 OMNIBUS survey of the Survey Research Centre at the University of Montreal. One section of this survey was aimed at measuring voters' intentions four weeks before the December

¹ Victor Tremblay, President, STATPLUS Statistical Consultants, PO Box 337, Ville Mont-Royal, Quebec H3P 3C6.

Table 1
Distribution of Voting Intentions
(with Non-Response)

	<i>n</i> ^a	%
Parti Québécois (PQ)	505	27.5
Quebec Liberal Party (QLP)	650	35.4
Other parties	62	3.4
Non-response	619	33.7
TOTAL	1,836	100.0

^a Number of weighted cases.

Table 2
Satisfaction with the Quebec Government and Voting Intentions
with Regard to the Provincial Election

Voting intentions	Satisfied (<i>n</i> = 555)	Dissatisfied (<i>n</i> = 656)
PQ	70.1%	17.3%
QLP	26.7%	76.1%
Other	3.2%	6.6%

1985 Quebec elections. The responses to the question regarding voting intentions given by the 1,836 individuals surveyed who intended to vote were distributed as in Table 1.

This table presents a situation where the response problem obviously cannot be ignored. Blindly distributing the non-responses in proportion to the other responses is a risky approach based on the supposition that those who did not express their voting intentions have the same profile as those who answered the question spontaneously.

The two consequences of such a high incidence of non-response are well known: potential bias and an increase in sampling error following effective reduction of sample size. Any adjustment technique must be aimed at reducing these two effects. When, as in this case, a high incidence of non-response can be foreseen, it is appropriate to include in the questionnaire correlated questions that can be used as a basis for eventual adjustments. For example, it may be very useful to ask the persons surveyed whether or not they are satisfied with the current government, given the close connection between this index and voting intentions, as shown in the following table.

As Table 2 shows, 70.1% of those satisfied with the government intended to support the party in power (the PQ). However, as might be expected, the situation was reversed among those who were dissatisfied: 76.1% of this number intended to vote for the QLP, which was the opposition party at the time.

Table 3

Satisfaction with the Government Cross-Classified with Whether Or Not an Answer Was Given to the Question Regarding Voting Intentions (Number of Weighted Cases)

	Satisfied	Dissatisfied	TOTAL
Answer given to question regarding voting intentions	$n_1 = 555$	$n_2 = 656$	$n = 1,211$
No answer given to question regarding voting intentions	236	334	570
TOTAL	$n'_1 = 791$	$n'_2 = 989$	$n' = 1,780^a$

^a This table excludes 56 nonresponses to the question on the satisfaction.

One of the techniques available for using this ancillary information is the creation of weighting classes based on satisfaction. Table 3 presents the complementary data required for making the adjustments.

If those who were satisfied and those who were dissatisfied are regarded as two weighting classes, statistical adjustment of the data takes the following form:

if p_{jc} = the proportion of respondents in class c who intend to support party j ;
 n_c = the number of persons in class c who answered the question regarding voting intentions;
 $n = \sum_c n_c$ = the size of subsample S_1 of those who answered the questions regarding voting intentions and satisfaction;
 n'_c = the total number of persons in class c ;
and $n' = \sum_c n'_c$ = the size of sample S of those who answered the question regarding satisfaction

The adjusted estimates of voting intentions are then calculated as follows:

$$p_j = (1/n) \sum_c n'_c p_{jc}.$$

This new estimate corresponds to introducing a corrective weight equal to $n'_c n / n_c n'$ for all respondents in class c .

This simple exercise illustrates the functioning of the well-known mechanism of statistical adjustment through construction of weighting classes based on traditional poststratification procedures. The questions which must be gone into in more depth for such an application are as follows:

1. What is the impact of this procedure on reduction of non-response bias?
2. How does this technique affect sampling error?
3. What are the best ancillary variables (or combinations or variables) for definition of the classes?
4. Up to what point is it advantageous to refine definition of the weighting classes?
5. What should be done with ancillary variables that also involve non-response?

To answer these questions properly, we must continue to develop the theoretical foundations for application of weighting classes.

3. IMPACT OF ADJUSTMENT PROCEDURE ON NON-RESPONSE BIAS

The most difficult challenge with respect to non-response is that of quantifying reduction of non-response bias following application of a given technique. If this challenge could be met, it would be possible to measure the bias and, consequently, to produce unbiased estimates.

However, we can still endeavour to understand more fully the mechanisms underlying non-response, in order to design instruments that would reduce as much as possible the impact of non-response on data quality.

One way of studying the problem is to consider it from the angle of response-probability theory, according to which we would stipulate that, for each unit U_i of the population, the probability of responding to the survey (or to a specific question asked) is α_i if that unit is selected. Even though this approach calls for the supposition that the α 's are not nil, the theory allows us to infer mathematical expressions for non-response bias with the application of a given method, in function of the observations X_i that we want to obtain and of the response probabilities α . This was the approach taken by Platek and Gray (1983); for estimating subtotal in weighting class c , by adjusting the sampling estimation using the inverse of the response rate in class c , they established that residual non-response bias could be expressed as follows:

$$B(\hat{X}_c) = \bar{\alpha}_c^{-1} \sum_{i=1}^{N_c} (\alpha_i - \bar{\alpha}_c) X_i \quad \text{where } \bar{\alpha}_c = N_c^{-1} \sum_{i=1}^{N_c} \alpha_i \quad (3.1)$$

and where N_c = the size of class c of the population.

Expression (3.1) reminds us that residual non-response bias exists following application of the correction factor only if, within class c , there is a correlation between the response probabilities and the characteristic measured.

Moreover, it is interesting to examine expression (3.1) in the special context of classification data—that is, where the $X_i = 0$ or 1. Using the notation introduced in the preceding section, it can be shown that the residual bias of \hat{X}_c following application of the correction factor can be written on the basis of expression (3.1) in the following form:

$$\begin{aligned} B(\hat{X}_c) &= N_c P_c \bar{\alpha}_c^{-1} (\bar{\alpha}_c^x - \bar{\alpha}_c) \\ &= N_c P_c (1 - P_c) \bar{\alpha}_c^{-1} (\bar{\alpha}_c^x - \bar{\alpha}_c^{\bar{x}}); \end{aligned}$$

where P_c = the real proportion of the units in class c that have characteristic X ;

$\bar{\alpha}_c^x$ = the average for response probabilities among the units in class c that have characteristic X ;

and $\bar{\alpha}_c^{\bar{x}}$ = the average for response probabilities among the units in class c that do not have characteristic X .

It is useful to reformulate $B(\hat{X}_c)$ as follows:

$$B(\hat{X}_c) = N_c \sigma_c^2 d_c(X, \bar{X})$$

where σ_c^2 = is the variance of characteristic X within class c

and $d_c(X, \bar{X}) = \bar{\alpha}_c^{-1} (\bar{\alpha}_c^x - \bar{\alpha}_c^{\bar{x}})$ is a standardized measurement of the distance between the average response probability for those who have characteristic X and that for those who do not have it within class c .

The non-response bias associated with estimation p' of P can therefore be expressed as:

$$\begin{aligned} B(p') &= B(N^{-1} \sum_c \hat{X}_c) \\ &= N^{-1} \sum_c N_c \sigma_c^2 d_c(X, \tilde{X}), \end{aligned} \tag{3.2}$$

Expression (3.2) provides a mathematical argument in support of the thesis frequently put forward that it is advantageous to construct categories that are as homogeneous as possible with respect to the phenomenon studied by partitioning the sample into segments, some of which tend to contain units with characteristic X , and some of which do not.

4. IMPACT OF ADJUSTMENT PROCEDURE ON SAMPLING ERROR

As you know, one of the consequences of the non-response problem is an increasing of random sampling error following reduction in the number of observations. It is revealing to examine to what extent the adjustment technique discussed here compensates for this loss of precision. A number of the authors referred to in the introduction have pointed out the potential danger in having corrective weights that are too large or too unstable, being based on a number of observations within a class that is too limited. Platek and Gray (1983) presented an approximate expression for the component of variance attributable to non-response following adjustment.

Although it is instructive regarding the general behaviour of this component of sampling variance, this mathematical development does not reveal the critical point beyond which refinement of the weighting classes adversely affects data accuracy.

In reality, we find ourselves in the following situation. The person conducting the survey has some reliable information with respect to a representative sample of the population being studied (for example, information regarding satisfaction with the government), but the data that interest him or her most for purposes of the survey (for example, information regarding voting intentions) are available only for a subsample, and he or she would like to use certain data from the base sample to improve the accuracy of the estimators. Whether we are talking about non-response at the level of the sampled units or about non-response at the level of specific questions in a questionnaire, the fundamental problem is the same. From the point of view of estimator variance, there is some analogy with double sampling, where data adjustment corresponds to application of the separate-ratio estimators—that is, to poststratification using categories definable on the basis of information available in the base sample. Of course, this analogy is unacceptable as far as analysis of the biasing effect of non-response is concerned, since one cannot support the hypothesis that the subsample of the respondents is probabilistically representative of the base sample. However, for purposes of studying estimator variance, the analogical approach is as useful as it is defensible.

More specifically, imagine the following situation. A simple random sample S of size n' gives us the distribution of a classification variable for the total population, with $\hat{N}'_c = (N/n')n_{c'}$ as the estimator of the number of units of the population belonging to class c . A simple random sample $S_1 \subset S$ of size $n=fn'$ ($0 < f < 1$) is chosen to measure the distribution of another classification variable X . For each of the units of S , we know the classification on the basis of the two variables described above.

We want to estimate the proportion P_j of units belonging to class j of variable X . The simple estimator inferred from S_1 is

$$p_j = (1/n) \sum_c n_c p_{jc}.$$

Moreover, the separate-ratio (poststratified) estimator can be expressed as follows:

$$p'_j = (1/n') \sum_c n'_c p_{jc}.$$

While all the units of sample S_1 are given a weight equal to 1 in expression p_j , we can see that, in expression for p'_j , the weight of the units varies, depending on the c class to which they belong. These "corrective" weights equal to n'_c/n use the complementary information available with respect to sample S as a whole for division into classes.

According to Tremblay (1975), if the formula for the variance of p_j is developed, keeping the terms to the size of the relative variance of the \hat{N}_c , the following is obtained:

$$\text{Var } p_j = \text{Var } p_j - [(1-f)/n] \left[\sum_c (P_{jc} - P_j)^2 P_c - \sum_c r_c P_{jc} (1 - P_{jc}) P_c \right] \quad (4.1)$$

where $r_c = N(1 - P_c)/nN_c$: the relative variance of the N_c estimator that is, $\hat{N}_c = (N/n)n_c$;

$P_{jc} = N_c/N$: the proportion of the population belonging to class c ;

$P_{jc} = Ep_{jc}$: the proportion of the units that have characteristic j in class c ;

$P_j = Ep_j$: the proportion of the units of the population that have characteristic j .

Equation (4.1) shows that the effectiveness of the technique of adjustment using weighting classes increases as interclass variance increases and, consequently, as intraclass variance decreases. It is easy to verify that, in the extreme case where there is maximal interclass variance – that is, where all of the P_{jc} are either 0 or 1:

$$\text{Var } p_j = P_j(1 - P_j)/n'.$$

that is, the variance that would have been obtained if all of the n' units had responded.

In addition, equation (4.1) reminds us that, in so far as the relative variances are negligible with respect to 1, it is advantageous to refine the partitioning, dividing the sample into a large number of classes. We thereby increase interclass variation and, by the same token, reduce the variance of p'_j .

However, refinement of the partitioning is limited by the presence of relative variances r_c . We should look at this situation a little more closely. Let us postulate that a first partitioning of the sample into a group of classes C' produces estimator p'_j as previously defined. Then let us postulate that a second, more refined partitioning C'' allows for the construction of estimator p''_j . If all of the classes coincide with the classes, except for one c class divided into two parts (c_1 and c_2 that is, $c = c_1 \cup c_2$), it would be interesting to find a simple criterion for determining which of the two partitions (C' or C'') produces the smallest variance, taking into account the r_c factors in expression (4.1) above. We know that:

$$\text{Var } p'' < \text{Var } p'$$

$$\begin{aligned} \Leftrightarrow G &= \sum_{c \in C''} (P_{jc} - P_j)^2 P_c - \sum_{c \in C'} (P_{jc} - P_j)^2 P_c \\ &> \sum_{c \in C''} r_c P_{jc} (1 - P_{jc}) P_c - \sum_{c \in C'} r_c P_{jc} (1 - P_{jc}) P_c = D. \end{aligned}$$

The left-hand member G of the inequality can be expressed thus:

$$\begin{aligned} G &= \sum_{c \in C''} P_{jc}^2 P_c - \sum_{c \in C'} P_{jc}^2 P_c \\ &= P_{jc_1}^2 P_{c_1} + P_{jc_2}^2 P_{c_2} - P_{jc}^2 P_c. \end{aligned} \quad (4.2)$$

If class c has been partitioned in the following way:

$$n_{c_1} = an_c \text{ and } n_{c_2} = (1-a)n_c \text{ when } 0 < a < 1$$

we know that $P_{jc} = aP_{jc_1} + (1-a)P_{jc_2}$, that $P_{c_1} = aP_c$, and, finally, that $P_{c_2} = (1-a)P_c$. Expression (4.2) can therefore be written compactly as follows:

$$G = P_c a(1-a) [P_{jc_1} - P_{jc_2}]^2.$$

Moreover, the right-hand member can be reduced to:

$$D = r_{c_1} P_{jc_1} (1 - P_{jc_1}) P_{c_1} + r_{c_2} P_{jc_2} (1 - P_{jc_2}) P_{c_2} - r_c P_{jc} (1 - P_{jc}) P_c.$$

With respect to the relevance of refining the partitioning, by replacing relative variances r_c with the expression previously established, noting that the terms P_{c_1} , P_{c_2} and P_c are negligible with respect to 1, we obtain:

$$D = (1/n) [P_{jc_1} (1 - P_{jc_1}) + P_{jc_2} (1 - P_{jc_2}) - P_{jc} (1 - P_{jc})]$$

Because of the convexity of function $P(1-P)$ and the fact that P is a linear combination P_{jc_1} and P_{jc_2} , the value of D is limited in an upwards direction by $1/4n$. Thus, for the variance of p'' to be smaller than that of p' , it is sufficient that:

$$P_c a(1-a) [P_{jc_1} - P_{jc_2}]^2 > 1/4n.$$

If P_c is estimated using n_c/n on the basis of subsample S_1 , this condition takes the following form:

$$DIF = |P_{jc_1} - P_{jc_2}| > \frac{1}{2} \sqrt{a(1-a)n_c} = DIFMIN \quad (4.3)$$

Inequality (4.3) therefore reveals a simple rule that is sufficient to make it advantageous to divide class c into $c_1 \cup c_2$. As we might have expected intuitively, the larger the number

Table 4.
Values of $DIFMIN = \frac{1}{2}\sqrt{a(1-a)}n_c$ (in %)

n	$a = 1/2$	$a = 1/4$	$a = 1/10$
1000	3.1%	3.7%	5.3%
400	5.0%	5.8%	8.3%
200	7.1%	8.2%	11.8%
100	10.0%	11.5%	16.7%
80	11.2%	12.9%	18.6%
60	12.9%	14.9%	21.5%
40	15.8%	18.3%	26.4%
20	22.4%	25.8%	37.3%
15	25.8%	29.8%	43.0%
10	31.6%	36.5%	52.7%

of respondents in class c (in sample S_1) or the greater the difference between the P_{jc_1} and P_{jc_2} proportions, the more advantageous it is to refine the partitioning of the classes. Table 4 above presents the minimal differences ($DIFMIN$) corresponding to various values of n_c and a .

The above table tells us that, for example, if we have a class containing 100 respondents which we are considering dividing into two more or less equal parts, there must be a difference of at least 10% between the two new classes where the j characteristic is concerned if the refinement of the partitioning is to help reduce sampling error. If there is less than a 10% difference between the two, refinement will serve no purpose, and may even increase the variability of the estimates produced. Moreover, we can see that if subclasses c_1 and c_2 are very unequal, the requirement regarding differentiated behaviour of their respondents with respect to characteristic j (that is P_{jc_1} vs P_{jc_2}) is stronger. Thus, if c_1 represents approximately 10% of c , the minimal difference ($DIFMIN$) is 16.7%.

In the specific case where class c is divided more or less equally between c_1 and c_2 , the minimal difference ($DIFMIN$) can be expressed very compactly:

$$DIFMIN = 1 / \sqrt{n_c}$$

In situations where class c is divided into several components ($c = c_1Uc_2U...Uc_k$), we can apply the test described here, considering the smallest of subclasses c_j on the one hand, and all of the rest on the other. Since, in this case, a (or $1-a$) may be small, we can simplify the rule expressed by inequality (4.3) and consider the minimal difference as follows:

$$DIFMIN = \frac{1}{2}\sqrt{\min_j (n_{c_j})}$$

It should be noted here that these results were developed by analogy in the context of sampling in two phases, and that the rules which have been arrived at may apply both to separate-ratio estimators and to poststratified estimators. For example, it is often useful to determine up to what point refinement of a poststratification produces more precise results. The rules set out here may therefore serve as a guide.

5. CRITERION FOR CHOOSING ADJUSTMENT VARIABLES

Looking once more at the survey of voters' intentions, we see that the degree of satisfaction with the government can certainly serve as an adjustment variable for non-response with respect to the question regarding voting intentions. However, is this really the best variable we could use? If the survey instrument contains other questions connected indirectly with voting intentions, on the basis of what criterion can we choose between, for example, satisfaction, certain sociodemographic profiles (language, education) and the perception as to who would make the best premier?

The two preceding sections show us that the more homogeneous the constructed classes are, the more variance of the adjusted estimates is reduced and the more likely it is that the bias itself will be smaller. It is therefore advantageous to create classes that maximize interclass variance of estimator p_j . With respect to algebraic expression (4.1), the partitioning chosen must maximize the quantity

$$INTERCL_j = \sum_c (P_{jc} - P_j)^2 P_c$$

For a multinomial variable X with parameters P_1, P_2, \dots, P_J , the problem is finding a statistic that incorporates all of the $INTERCL$ quantities ($j = 1, \dots, J$). In this case, χ^2 merits consideration, since

$$\chi^2 = N \sum_j \sum_c (P_{jc} - P_j)^2 P_c / P_j = N \sum_j P_j [INTERCL_j / P_j^2]$$

In other words, χ^2 is equal to a linear combination of the relative values of the $INTERCL_j$'s weighted in function of the P_j 's. On the other hand, since $B_j = INTERCL_j / P_j (1 - P_j)$ measures the proportion of the variance explained by division into classes, there is also justification for considering the statistic

$$\sum B_j = \sum_j \sum_c (P_{jc} - P_j)^2 P_c / P_j (1 - P_j).$$

Note that the latter statistic is equivalent to χ^2 in three specific situations: a) when X is dichotomous; b) when the P_j 's are almost equal; and c) when the P_j 's are all small. In the multinomial case, where it is important to refine estimation of a P for a particular j index, we can therefore dichotomize variable X in function of this j index and use χ^2 as a performance criterion for division into classes. For our example, we will use χ^2 , since this statistic is produced directly by most of the software used for processing survey data.

6. APPLICATION AND INHERENT PROBLEMS

In the preceding discussion, we found a criterion for evaluating the performance of weighting classes. In practice, however, variables which best explain variance may also be affected by the non-response problem. This complicates the choosing of weighting classes to some extent.

The following table presents a list of variables deemed interesting a priori by a researcher for the purpose of weighting to adjust for non-response with respect to the voting-intentions

question. For each potentially useful variable, there is a description of the value of χ^2 , the number of missing values and the total number of missing values when the variable is cross-ed with the question on voting intentions. Remember that the latter question, taken alone, accounted for 619 non-responses in the survey.

The value of χ^2 is very revealing with respect to the predictive force of the different variables involved. For example, we can see that, among the sociodemographic variables, only mother tongue has an impact that merits attention. On the other hand, some thematic questions show an unequivocal link with voting intentions – in particular, that regarding degree of satisfaction with the present government and that which asks which of the two main party leaders would be the best premier. It is clear that the more a question is perceived as being connected with the basic question, the more difficult it is to obtain responses. Only 56 non-responses were recorded for the more insignificant question regarding satisfaction with the government (approximately 3% of the sample), but there were 392 non-responses when people were asked who would be the best premier!

In the creation of weighting classes, it is therefore advantageous to try to use variables strongly correlated with the phenomenon being studied, as well as variables which are both strongly correlated and characterized by an excellent response rate. In addition, by crossing the relevant variables with each other, we can create classes that are more homogeneous and, consequently, increase the value of χ^2 . Obviously, the degree of refinement of the classes must be in line with the limiting criterion previously expressed by equation (4.3).

Table 5

List of Variables That Might Be Useful for Compensation, through Weighting, for the Effect of Non-Response with Respect to the Question on Voting Intentions

Variable ^a	Value of χ^2	Number of missing data on the variable	Number of missing data upon cross-classification with voting intentions
Age (6)	34	4	620
Education (4)	8	3	621
Mother tongue (2)	96	0	619
Degree of satisfaction with Quebec government (4)	382	56	625
Degree of satisfaction with Quebec government (2)	346	56	625
Identification of best premier (3)	773	392	686
Vote in 1981 provincial election	109	269	658
Interest in politic	1	1	619
Degree of satisfaction with federal government (4)	39	58	631
Voting intentions at federal level (4)	288	694	832

^a The figures in parentheses indicate the number of classes considered for the variables in question.

Consider, for example, the formation of weighting classes on the basis of three variables that are explanatory with respect to voting intentions – namely, identification of the party leader who would be the best premier (3 response categories), degree of satisfaction with the government (4 categories) and mother tongue (2 categories). At this stage, the idea is to project the voting intentions determined for the respondents in a given class onto all of the individuals in that class – that is, those for whom it has been possible to establish a classification. The first step in the process is to refine the classes as much as possible on the basis of the three variables involved and produce a cross tabulation of voting intentions in accordance with these twenty-four (3x4x2) classes. Referring either to the criterion revealed in equation (4.3) or to Table 4, we eliminate through combination those classes which are too small. Where necessary, we therefore group together “similar” classes – that is, classes that have a similar voting-intentions profile. We are then in a position to produce a table like that on the following page, in which voting intentions are cross-classified following this new division. An examination of the data may also suggest a few groupings. In addition, Table 6 presents other relevant data. For example, the last two lines compare by class the number of individuals who answered the question regarding voting intentions with the total number of individuals surveyed who can be classified in accordance with the three variables involved. From this, we obtain a first weighting system. In the example, there are 283 persons overall who can be classified, but whose voting intentions are not known. In addition, the overall value of χ^2 is 891, a distinct improvement over the situation when the variables were taken alone (Table 5).

Finally, in the B_j column, for each P_j , there appear estimates of the percentage of the variance that can be attributed to interclass variance. These B_j 's measure the increase in precision (variance reduction) that can be attributed to adjustment of the data in accordance with the type of partition chosen. This is clear if we rewrite equation (4.1) as follows (disregarding relative variances):

$$\text{Var } p'_j = \text{Var } p_j - (1 - f)B_j \text{Var } p_j$$

Having a B_j equal to 61.9% for estimation of intention to vote for the PQ means that, from the point of view of variance reduction, adjustment of the data is equivalent to having recuperated in the field 61.9% of the 283 non-responses for the question on voting intentions.

We now have the residual problem of determining how to adjust for non-response for specific questions using variables that have themselves been affected by non-response.

In the example produced through division in accordance with Table 6, it is clear that a significant portion of the non-responses with respect to voting intentions is not corrected through this kind of weighting. In effect, we are left with 409 cases of non-response that cannot be dealt with in this fashion, since classification with respect to a reference variable cannot be determined. One possibility that might be explored here is establishment of a weighting system that would allow us to use, for each non-respondent, the maximum number of variables available for estimating the missing data. For example, the voting-intentions profile of persons who did not respond to the question on voting intentions or to that asking who would be the best premier, but who we know are Francophone and are satisfied with the government in power, would be inferred on the basis of the voting-intentions profile of the Francophone respondents satisfied with the government. A weighting system can easily be developed for this process of attribution.

Table 6.
Study of a Partitioning of the Sample

Best premier		Johnson (PQ)			
Satisfied government	Very	Fairly	Very or Fairly	Not very or not all all	
Mother tongue	Franco-phone	Franco-phone	Non-franco-phone	Franco-phone	Non-franco-phone
% vote PQ	100	88.3	42.7	63.4	32.9
% vote PLQ	0.0	9.5	45.8	30.2	61.4
% vote Other	0.0	2.2	11.5	6.4	5.7
Number of respondents for the classification and voting-intentions questions	51	342	37	133	22
Number of respondents for the classification questions	59	404	50	203	28
Best premier		Bourassa (PLQ)			
Satisfied government	Very or fairly	Not very			Not at all
Mother tongue	Franco-phone	Non-franco-phone	Franco-phone	Non-franco-phone	Franco-phone
% vote PQ	12.2	0.0	6.5	1.2	3.5
% vote PLQ	86.5	85.9	92.4	98.8	93.9
% vote Other	1.3	14.1	1.1	0.0	2.6
Number of respondents for the classification and voting-intentions questions	64	21	156	49	159
Number of respondents for the classification questions	81	24	178	54	175
Best premier	Other than Johnson	Neither Bourassa nor Johnson		TOTAL	B_j (%)
Satisfied government	Not at all	Very or fairly	Not very or not all all		
Mother tongue	Non-franco-phone	----	----		
% vote PQ	0.0	14.3	4.6	42.7	61.9
% vote PLQ	100.0	73.5	54.9	52.6	58.7
% vote Other	0.0	12.2	40.5	4.7	15.1
Number of respondents for the classification and voting-intentions questions	42	17	51	1144	($\chi^2 = 891$)
Number of respondents for the classification questions	49	32	89	1427	

REFERENCES

- BAILAR, B.A., BAILEY, L., and CORBY, C. (1978). A comparison of some adjustment and weighting procedures for survey data. In *Survey Sampling and Measurement* (Ed. N. Krishnan Namboodiri), New York: Academic Press, 175-198.
- CHAPMAN, D.W. (1983). An investigation of nonresponse imputation procedures for the health and nutrition examination survey. In *Incomplete Data in Sample Surveys*, Volume 1 – Report and Cases Studies (Eds. W.G. Madown, H. Nisselson, and I. Olkin), New York: Academic Press, 435-483.
- KISH, L. (1978). On the future of survey sampling. In *Survey Sampling and Measurement* (Ed. N. Krishnan Namboodiri), New York: Academic Press, 13-21.
- OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*, Volume 2 – Theory and Bibliographies (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 143-184.
- PANEL ON INCOMPLETE DATA (1983). Part I – Report. In *Incomplete Data in Sample Surveys*. Volume 1 – Report and Cases Studies (Eds. W.G. Madow, H. Nisselson, and I. Olkin), New York: Academic Press, 3-103.
- PLATEK, R., and GRAY, G.B. (1983). Part V – Imputation Methodology: Total Survey Error. In *Incomplete Data in Sample Surveys*, Volume 2 – Theory and Bibliographies (Ed. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 249-333.
- TREMBLAY, V. (1975). On the improvement of sample surveys estimates. *Survey Methodology*, 1, 181-196.

A Study of the Effects of Imputation Groups in the Nearest Neighbour Imputation Method for the National Farm Survey

SIMON CHEUNG and CRAIG SEKO¹

ABSTRACT

A new processing system using the nearest neighbour (N-N) imputation method is being implemented for the National Farm Survey (NFS). An empirical study was conducted to determine if the NFS estimates would be affected by using imputation groups based on type of farm. For the specific imputation rule examined, the study showed evidence that the effect might be small.

KEY WORDS: National Farm Survey; Item non-response; Nearest neighbour imputation; Match variable transformation.

1. INTRODUCTION

The National Farm Survey (NFS) is an annual multi-purpose survey of agricultural activity in Canada. The survey uses a 2-frame sample design i.e. a list frame of large farms (based on the quinquennial Census of Agriculture) and an area frame of agricultural land. The largest units in the list frame are sampled with certainty (i.e. with probability one) because of their disproportionate impact on the survey estimates. These units are called specified farms. The remaining farms in the list frame are stratified and sampled. The small farms in the survey population, which are comparatively very large in number, are covered by the area frame and sampled less extensively than the list frame farms. Thus three samples are selected: specified, list and area. The detailed NFS sample design has been described by Davidson and Ingram (1983), and Davidson (1984).

The NFS is processed by a system adopted from predecessor surveys. This system employs the sequential hot-deck imputation method to adjust for unit and item non-response (Philips 1979). A new survey processing system will be implemented in 1987 in order to integrate all the agricultural surveys conducted by Statistics Canada. This system will use the nearest neighbour (N-N) imputation method to adjust for item non-response. The decision to implement the N-N imputation method was based on many reasons, among which there are three important ones: First, the use of the N-N method is theoretically more justified than the exact-matching sequential hot-deck method since the survey collects mostly quantitative data. Second, empirical studies, e.g. Kovar (1982), suggest that the two imputation methods would yield similar estimates for the NFS with the N-N method resulting in fewer outliers i.e. imputed data which have disproportionate contributions to the survey estimates. Third, switching to this new imputation method for the NFS would help standardize the survey methodology of all agricultural surveys, a long term goal of Statistics Canada. Currently, the Census of Agriculture and the Farm Tax Data Survey both use the N-N imputation methodology.

This paper reports on an empirical study which attempts to provide information that will help in a more efficient implementation of the new imputation method. The next section describes briefly the N-N imputation method adopted in our study. Section three presents the study procedure and the main results obtained. Finally, we discuss our preliminary observations drawn from the results in section four.

¹ Simon Cheung and Craig Seko, Business Survey Methods Division, Statistics Canada, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

2. NEAREST NEIGHBOUR IMPUTATION METHOD

The method of donor imputation, in general, is to replace the missing or invalid values of a respondent (recipient) with the valid response of another respondent (donor) who is deemed to have the same characteristics as the recipient. The sequential hot-deck imputation method identifies donors sequentially in the course of processing as those reporting the same values as the recipient in the pre-specified match variables. This method, however, often fails to obtain an exact match when a match variable assumes a large number of possible values. To alleviate this, the range of the match variable is split into intervals and the donor is obtained by matching on the interval code. In nearest neighbour imputation, this problem is solved by selecting a donor based on a multivariate distance measure which represents the degree of similarity between the donor and the recipient as defined by the pre-specified match variables. The more similar two respondents are with respect to the match variables, the smaller the magnitude of the distance. Thus, the best donor for a recipient is the donor candidate which has the smallest distance value from the recipient, i.e. its nearest neighbour in the sense of statistical distance.

The nearest neighbour imputation method used in this study was proposed by Sande (1976, 1981). This method uses the maximum norm based on transformed data as the distance function. The method is described briefly below.

Let $X = (x_1, x_2, x_3, \dots, x_k)$ be a vector of k match variables. Each match variable x_j is transformed by $t_j = \hat{F}(y)$, where $\hat{F}(y)$ is the empirical distribution function of x_j . Note that t_j follows the uniform distribution over $[0, 1]$. Then the distance between a given recipient X^r and a donor candidate X^d defined by the maximum norm is

$$d(X^r, X^d) = \max_j |t_j^r - t_j^d|,$$

where t_j^r and t_j^d are the transformed values of the j^{th} match variable x_j in X^r and X^d , respectively. The donor candidate with the smallest d-value will be selected and its response will be copied for the missing item of the recipient. The uniform transformation may be considered as an objective method to scale the match variables regardless of their natural distributions.

3. EMPIRICAL STUDY

3.1 Motivation

In adopting the nearest neighbour imputation method for the NFS, some issues regarding detailed implementation of this method need to be resolved, particularly in regards to transforming match variables. The method of uniform transformation in the N-N imputation could be applied using all the records in the sample or using only subsets of the sample data. A group of unit respondents in which imputation for non-response takes place is called an imputation group. Different imputation groups would yield different transformed values which in turn would result in different selection of donor records.

It was conjectured that transforming match variables within an imputation group defined by a homogeneity criterion which is closely related to the item to be imputed would result in a more correct scaling of the match variables, and hence would yield better imputed data. For example, in the NFS one may expect that match variable transformation within imputation groups defined by farm type should yield better imputed data and hence better estimates, 'better' being in the sense of bias and variance reduction. Unfortunately, the transformation of match variables is costly in terms of computer resources. If one does not need to transform within homogeneous imputation groups, savings in computer costs can be realized.

The main objective of the study was to answer the following question in an experimental setting: 'Do the two methods of match variable transformation, i.e., transformation using all records vs. within farm type groups, yield substantially different survey estimates? If so, which method yields better estimates?'

3.2 Data Used in the Study

After consultation with the subject matter analysts, the 1984 NFS sample for the province of Alberta was selected for the study. The sample of approximately 2000 farms consists of 50% crop farms, 27% livestock farms and 23% mixed farms. The population percentages of the three farm types were estimated to be 52%, 27% and 21% respectively. Farm types were assigned according to the main source of projected agricultural receipts of a farm. If at least 75% of a farm's projected agricultural receipts came from its livestock inventory, the farm was classified as a livestock farm. A similar rule was used to classify crop farms. The remaining farms were classified as mixed farms.

3.3 Method of the Study

We assumed that the data was 'clean', even though it contained imputed values via the sequential hot-deck imputation procedure. Once the data had been classified by farm type, the following procedure was followed:

- i) Ten per cent of the values for each imputation variable was randomly set to a missing value within each farm type. This error generation was done independently for each imputation variable.
- ii) The generated non-responses were imputed using the N-N imputation method based on the two sets of imputation groups defined by the whole sample (called 'whole') and by farm type (called 'by-type'). The imputation procedures were carried out using the Numerical Edit/Imputation System (Statistics Canada 1982), as implemented within the P-STAT statistical package (Buhler and Buhler 1978).
- iii) The NFS weighted estimates for the variable totals for the province and for each farm type were produced based on each set of imputed data.
- iv) These steps were repeated 10 times to get 10 independent replications (i.e., simulations), and the results were averaged over the ten replications for each imputation variable. This average estimate was then compared with the estimate obtained based on the 'clean' file, both at the provincial level and for each farm type.

The whole experiment was repeated for higher non-response rates of 15% and 20% in order to observe the impact of nonresponse rates.

The imputation and match variables used in the study are shown below:

Imputation Variables

- UTIL = Utility expenses
- AUTO = Farm vehicle and machinery operating expenses
- TAX = Property tax

Match Variables

Farm type (exact matching)

FEED = Feed expense

SEED = Seed expense

INCOME = Gross agricultural receipts

In addition, the donor's sample type was restricted by the recipient's. Recall that three types of samples are used in the NFS: specified, list, and area. A specified farm can be imputed by a farm from any of the sample types but can not be a donor to a list or area farm. Similarly, a farm from the list sample can be imputed from a farm in either the list or area samples but can only be a donor to farms that are in the list sample or are specified. Finally, farms in the area sample can only be imputed by another area farm but can serve as a donor to any of the three samples. These restrictions arise from the premise that if a list or specified farm was allowed to impute for an area farm, the imputed value could potentially raise the survey estimates to an unacceptable level because of the higher sampling weights associated with area farms.

3.4 The Empirical Distribution Functions of the Match Variable

Figure 1 shows the unweighted empirical distribution functions of the three match variables which are obtained from the imputation groups defined by the whole sample and by farm type. Note that the differences are substantial and hence could lead to the selection of different donor records for a given recipient.

3.5 Results

The results are tabulated in Table 1. For each imputation variable (UTIL, AUTO or TAX), each of the two sets of imputation groups (whole vs. by-type), and each level of non-response rate (10%, 15% or 20%), the average value of the ten estimates for the variable total was calculated over the ten replications. The bias of this average value is displayed as a percentage of the "clean" estimate. The average cv over the ten replicates is also displayed as a percentage.

4. OBSERVATIONS AND DISCUSSION

This study imputed for three farming expense variables. The donor records were selected by exact matching on farm type and by nearest-neighbour matching on three variables: gross agricultural receipts, feed expense and seed expense. The two expense match variables were believed to be of different effectiveness for the three farm types. For example, feed expense was expected to work better for livestock farms but not so for crop farms, etc. The strength of correlation between the match variables and the imputation variables presented in Table 2 seems to support this expectation.

Therefore the homogeneous subsets based on type of farm have differing relationships for the match variables. This might imply that transformations using imputation groups defined by these subsets would perform better than using the entire sample as an imputation group. The results, however, indicate that using these homogeneous subsets as imputation groups does not seem to yield substantially different estimates or lower bias. The bias itself

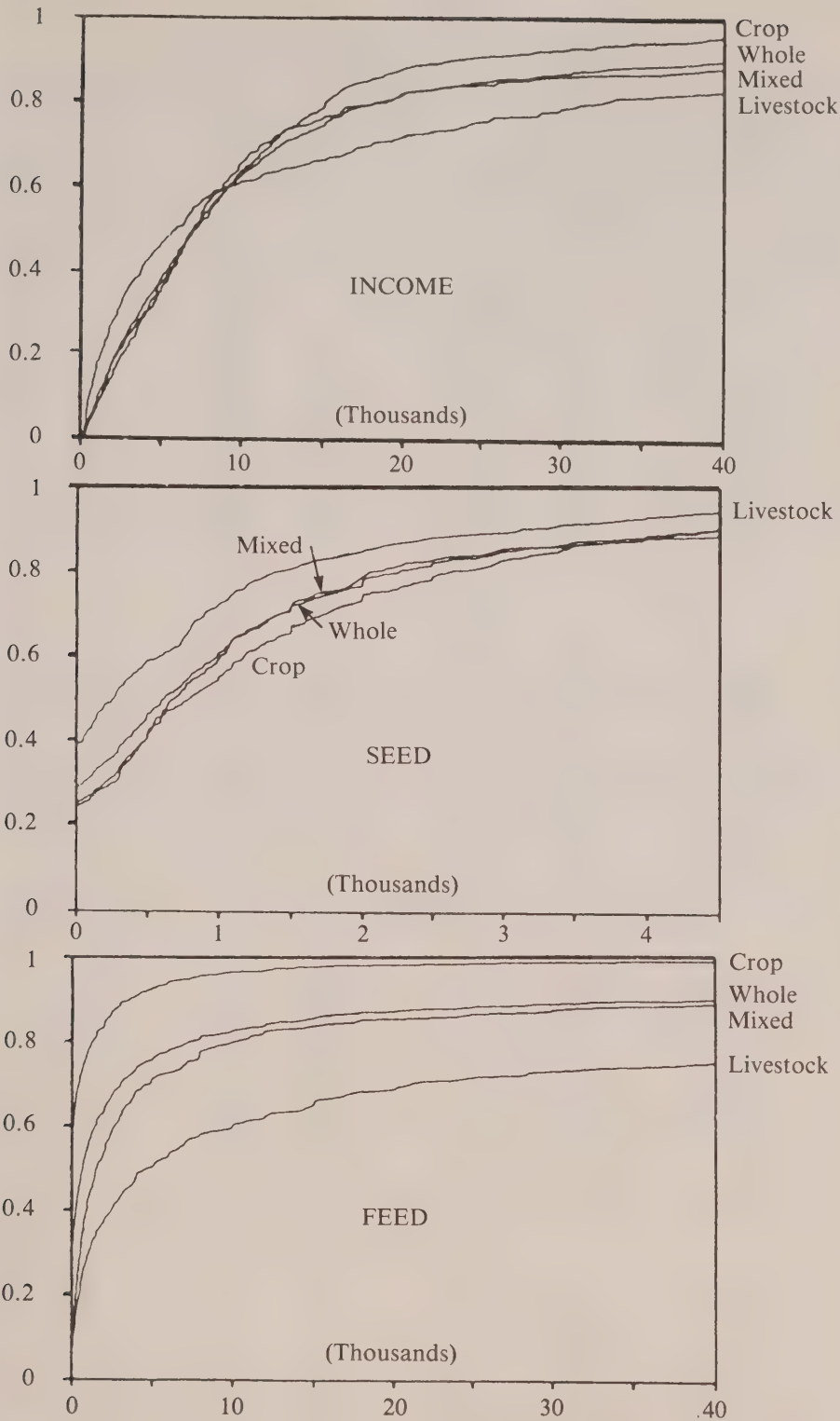


Figure 1: Empirical Distribution Functions of Match Variables

Table 1
Percentage Bias and cv's for the Totals of the Imputation
Variables after Imputation

Non-response rate	Imputation group	Imputation Variables					
		UTIL		AUTO		TAX	
		% Bias	% cv	% Bias	% cv	% Bias	% cv
All Farms in Sample							
clean			3.137		2.831		3.224
10%	by-type	0.176	3.165	-0.004	2.849	0.228	3.260
	whole	0.124	3.143	-0.074	2.840	0.199	3.296
15%	by-type	0.339	3.195	0.604	2.885	0.255	3.275
	whole	0.336	3.131	0.278	2.870	-0.624	3.289
20%	by-type	0.869	3.173	0.023	2.875	-0.715	3.280
	whole	0.554	3.111	-0.150	2.843	-0.877	3.285
Crop Farms							
clean			4.829		4.092		4.536
10%	bt-type	0.023	4.872	0.516	4.159	0.200	4.574
	whole	-0.221	4.829	0.328	4.155	0.371	4.625
15%	by-type	0.468	4.981	0.611	4.200	0.855	4.695
	whole	0.156	4.863	-0.199	4.231	-0.026	4.672
20%	by-type	0.402	5.008	0.620	4.238	-1.201	4.770
	whole	-0.170	4.944	0.129	4.227	-1.158	4.699
Livestock Farms							
clean			6.770		5.596		9.527
10%	by-type	0.125	6.798	-0.885	5.575	0.688	9.471
	whole	0.687	6.800	-0.487	5.532	-0.093	9.515
15%	by-type	0.234	6.829	0.156	5.523	0.346	9.325
	whole	0.789	6.797	0.646	5.533	-1.666	9.227
20%	by-type	1.526	6.920	-0.370	5.538	0.654	9.250
	whole	1.136	6.830	-0.051	5.495	-0.354	9.565
Mixed Farms							
clean			7.433		7.190		6.993
10%	by-type	0.570	7.519	-0.549	7.175	-0.092	7.029
	whole	0.093	7.507	-0.715	7.132	-0.009	7.027
15%	by-type	0.219	7.404	0.957	7.150	-1.437	7.143
	whole	0.115	7.407	1.142	7.107	-1.335	7.152
20%	by-type	0.984	7.541	-1.108	6.984	-0.599	7.010
	whole	1.303	7.595	-0.927	7.001	-0.576	7.050

Table 2
Correlation Coefficients between Match and
Imputation Variables^a

Farm Type	Imputation variable	Match variables		
		FEED	SEED	INCOME
whole	UTIL	0.46	0.39	0.50
	AUTO	0.34	0.18	0.50
	TAX	0.10	0.16	0.27
crop	UTIL	0.13	0.57	0.69
	AUTO	0.25	0.28	0.65
	TAX	0.18	0.19	0.48
livestock	UTIL	0.64	0.25	0.51
	AUTO	0.41	0.47	0.52
	TAX	0.13	0.25	0.28
mixed	UTIL	0.55	0.49	0.76
	AUTO	0.48	0.46	0.73
	TAX	0.24	0.45	0.55

^a The coefficients are based on unweighted data from the 1984 NFS core sample in Alberta.

seems negligible at low rates of non-response. As the non-response rate rises, the bias grows but is still not substantial. Except for the variable TAX, the differences between the estimates seldom exceed the 95% confidence limits. In the case of TAX, statistical significance, when detected, is usually at the 15% and 20% non-response rates. Unfortunately, the average estimates for the variables UTIL and TAX do show a pattern of consistent, positive bias. No explanation is obvious for this observation and further investigation is warranted to uncover the potential source of bias.

Thus, there is no need to transform match variables by imputation groups defined by farm type for the imputation studied; transforming match variables using the whole sample leads to very similar survey estimates. This may not be the case for other imputation rules and patterns of non-response that are not random. These are topics for future studies. Although the imputed estimates compare well with the clean estimates in practical terms, however, there may still be some unknown sources of bias. These sources, if they exist, may be related to this imputation method, to the imputation rule examined in this study or some other unidentified factor. It is suggested that the presence of bias be confirmed and if confirmed, its source determined. Further study is recommended to this end as well as to aid in determining future imputation rules for the National Farm Survey.

5. ACKNOWLEDGEMENT

The authors would like to thank Nanjamma Chinnappa and Allison Miller for their suggestions and support during the course of the study. We are also thankful for the comments of the referee of this paper.

REFERENCES

- BUHLER, S. and BUHLER, R. (1978). *P-Stat 78 Users's Manual*. P-Stat Inc., Princeton, N. J., U. S. A.
- DAVIDSON, G. (1984). 1983 National Farm Survey. Note on the sample design and estimation procedures. Working Paper, Institution and Agriculture Survey Methods Division, Statistics Canada.
- DAVIDSON, G., and INGRAM, S. (1983). Methods used in designing the National Farm Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 220-225.
- KOVAR, J. (1982). A closer look at the nearest neighbour/hot deck imputation methods: An empirical study. Working Paper, Institution and Agriculture Survey Methods Division, Statistics Canada.
- PHILIPS, J. (1979). Imputation techniques used for the F.E.S. Working Paper. Institution and Agriculture Survey Methods Division, Statistics Canada.
- SANDE, G. (1976). Searching for numerically matched records. Unpublished manuscript, Business Survey Methods Division, Statistics Canada.
- SANDE, G. (1981). Descriptive statistics used in monitoring edit and imputation process. *Proceedings of the 13th Symposium on the Interface*. Pittsburgh, Pennsylvania.
- STATISTICS CANADA. (1982). *The Numerical Edit and Imputation Subsystem for P-Stat - A User's Guide*. Special Resources Subdivision, Systems Development Division, Statistics Canada.

Il n'y a donc pas besoin de transformer les variables d'appariement par groupes d'imputation définis selon le type de ferme pour l'imputation étudiée; la transformation des variables d'appariement utilisant l'ensemble de l'échantillon donne des estimations d'enquête très semblables. Ceci pourrait ne pas être le cas pour d'autres règles et régimes d'imputation de non-réponse non aléatoires. Ceci devrait faire l'objet d'autres études. Les estimations imputées se comparent favorablement aux estimations propres en termes pratiques. Cependant, des causes inconnues de biais peuvent encore exister. Ces causes, si elles existent, peuvent se rattacher à cette méthode d'imputation, à la règle d'imputation examinée dans la présente communication ou à un autre facteur non identifié. On suggère que la présence de biais soit confirmée, et si c'est le cas, de déterminer sa cause. D'autres études sont recommandées pour cela, ainsi que pour aider à déterminer les règles d'imputation futures pour l'enquête nationale sur les fermes.

5. REMERCIEMENTS

Les auteurs voudraient remercier Nanjamma Chinnappa et Allison Miller pour leurs suggestions et leur soutien pendant la préparation de la présente communication. Nous voulons également remercier le critique de la communication pour ses commentaires utiles.

BIBLIOGRAPHIE

BUHLER, S. and BUHLER, R. (1978). *P-Stat 78 Users' Manual*. P-Stat Inc., Princeton, N. J., U. S. A.

DAVIDSON, G. (1984). 1983 National Farm Survey. Note on the sample design and estimation procedures. Document de travail, Division des méthodes d'enquêtes-institutions et agriculture, Statistique Canada.

DAVIDSON, G., and INGRAM, S. (1983). Methods used in designing the National Farm Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 220-225.

KOVAR, J. (1982). A closer look at the nearest neighbour/hot deck imputation methods: An empirical study. Document de travail, Division des méthodes d'enquêtes-institutions et agriculture, Statistique Canada.

PHILIPS, J. (1979). Imputation techniques used for the F.E.S. Document de travail, Division des méthodes d'enquêtes-institutions et agriculture, Statistique Canada.

SANDE, G. (1976). Searching for numerically matched records. Document non-publié, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

SANDE, G. (1981). Descriptive statistics used in monitoring edit and imputation process. *Proceedings of the 13th Symposium on the Interface*. Pittsburgh, Pennsylvania.

STATISTICS CANADA. (1982). *The Numerical Edit and Imputation Subsystem for P-Stat - A User's Guide*. Sous-division des ressources spéciales, Division du développement informatique, Statistique Canada.

Tableau 2

Coefficients de corrélation entre les variables d'appariement et d'imputation^a

Type de ferme	Variables d'imputation	PROVENDES	SEMENCES	REVENU
Variables d'appariement				

ensemble	UTIL	0.46	0.39	0.50
	AUTO	0.34	0.18	0.50
	TAX	0.10	0.16	0.27
cultures	UTIL	0.13	0.57	0.69
	AUTO	0.25	0.28	0.65
	TAX	0.18	0.19	0.48
bétail	UTIL	0.64	0.25	0.51
	AUTO	0.41	0.47	0.52
	TAX	0.13	0.25	0.28
mixte	UTIL	0.55	0.49	0.76
	AUTO	0.48	0.46	0.73
	TAX	0.24	0.45	0.55

^a Les coefficients sont basés sur des données non pondérées provenant de l'échantillon central de l'ENF de 1984 pour l'Alberta.

4. OBSERVATIONS ET DISCUSSION

Cette étude a imputé trois variables de dépenses agricoles. Les enregistrements donneurs ont été sélectionnés par un appariement exact selon le type de ferme et l'appariement du plus proche voisin sur trois variables: les recettes agricoles brutes, les dépenses en provendes et les dépenses en semences. Les deux variables d'appariement des dépenses ont été considérées comme ayant une efficacité différente pour les trois types de ferme. Ainsi, on s'attendait à ce que les dépenses en provendes donnent de meilleurs résultats pour les fermes de bétail et non pas pour les fermes de cultures, etc. La force de la corrélation entre les variables d'appariement et les variables d'imputation présentée au tableau 2 semble confirmer cette attente. Par conséquent, les sous-ensembles homogènes basés sur le type de ferme ont des relations différentes pour les variables d'appariement. Ceci pourrait signifier que les transferts utilisant les groupes d'imputation définis par ces sous-ensembles donneraient de meilleurs résultats que si l'on employait tout l'échantillon comme groupe d'imputation. Les résultats, cependant, indiquent que ces sous-ensembles homogènes utilisés comme groupe d'imputation ne semblent pas donner des estimations sensiblement différentes ou un biais plus bas. Ce dernier semble négligeable à des taux de non-réponse bas. À mesure que les taux de non-réponse augmentent, le biais augmente, mais n'est pas appréciable. À l'exception de la variable TAX, les différences entre les estimations dépassaient rarement les limites de l'intervalle de confiance à 95%. Dans le cas de TAX, la signification statistique, lorsqu'elle a été détectée, s'observe habituellement pour les taux de non-réponse de 15% et 20%. Malheureusement, les estimations moyennes des variables UTIL et TAX se caractérisent par un biais systématique et positif. Il n'y a aucune explication pour cette observation et d'autres études sont nécessaires pour découvrir la cause possible de ce biais.

Tableau 1
Biais et CV en pourcentage pour les totaux des variables d'imputation
après imputation

Taux de non-réponse	Groupe d'imputation	Variables d'imputation					
		UTIL		AUTO		TAX	
		Biais en %	cv en %	Biais en %	cv en %	Biais en %	cv en %

propre		3.137		2.831		3.224	
10%	selon le type	0.176	3.165	-0.004	2.849	0.228	3.260
10%	intégral	0.124	3.143	-0.074	2.840	0.199	3.296
15%	selon le type	0.339	3.195	0.604	2.885	0.255	3.275
15%	intégral	0.336	3.131	0.278	2.870	-0.624	3.289
20%	selon le type	0.869	3.173	0.023	2.875	-0.715	3.280
20%	intégral	0.554	3.111	-0.150	2.843	-0.877	3.285

Fermes de cultures

propre		4.829		4.092		4.536	
10%	selon le type	0.023	4.872	0.516	4.159	0.200	4.574
10%	intégral	-0.221	4.829	0.328	4.155	0.371	4.625
15%	selon le type	0.468	4.981	0.611	4.200	0.855	4.695
15%	intégral	0.156	4.863	-0.199	4.231	-0.026	4.672
20%	selon le type	0.402	5.008	0.620	4.238	-1.201	4.770
20%	intégral	-0.170	4.944	0.129	4.227	-1.158	4.699

Fermes de bétail

propre		6.777		5.596		9.527	
10%	selon le type	0.125	6.798	-0.885	5.575	0.688	9.471
10%	intégral	0.687	6.800	-0.487	5.532	-0.093	9.515
15%	selon le type	0.234	6.829	0.156	5.523	0.346	9.325
15%	intégral	0.789	6.797	0.646	5.533	-1.666	9.227
20%	selon le type	1.526	6.920	-0.370	5.538	0.654	9.250
20%	intégral	1.136	6.830	-0.051	5.495	-0.354	9.565

Fermes mixtes

propre		7.433		7.190		6.993	
10%	selon le type	0.570	7.519	-0.549	7.175	-0.092	7.029
10%	intégral	0.093	7.507	-0.715	7.132	-0.009	7.027
15%	selon le type	0.219	7.404	0.957	7.150	-1.437	7.143
15%	intégral	0.115	7.407	1.142	7.107	-1.335	7.152
20%	selon le type	0.984	7.541	-1.108	6.984	-0.599	7.010
20%	intégral	1.303	7.595	-0.927	7.001	-0.576	7.050

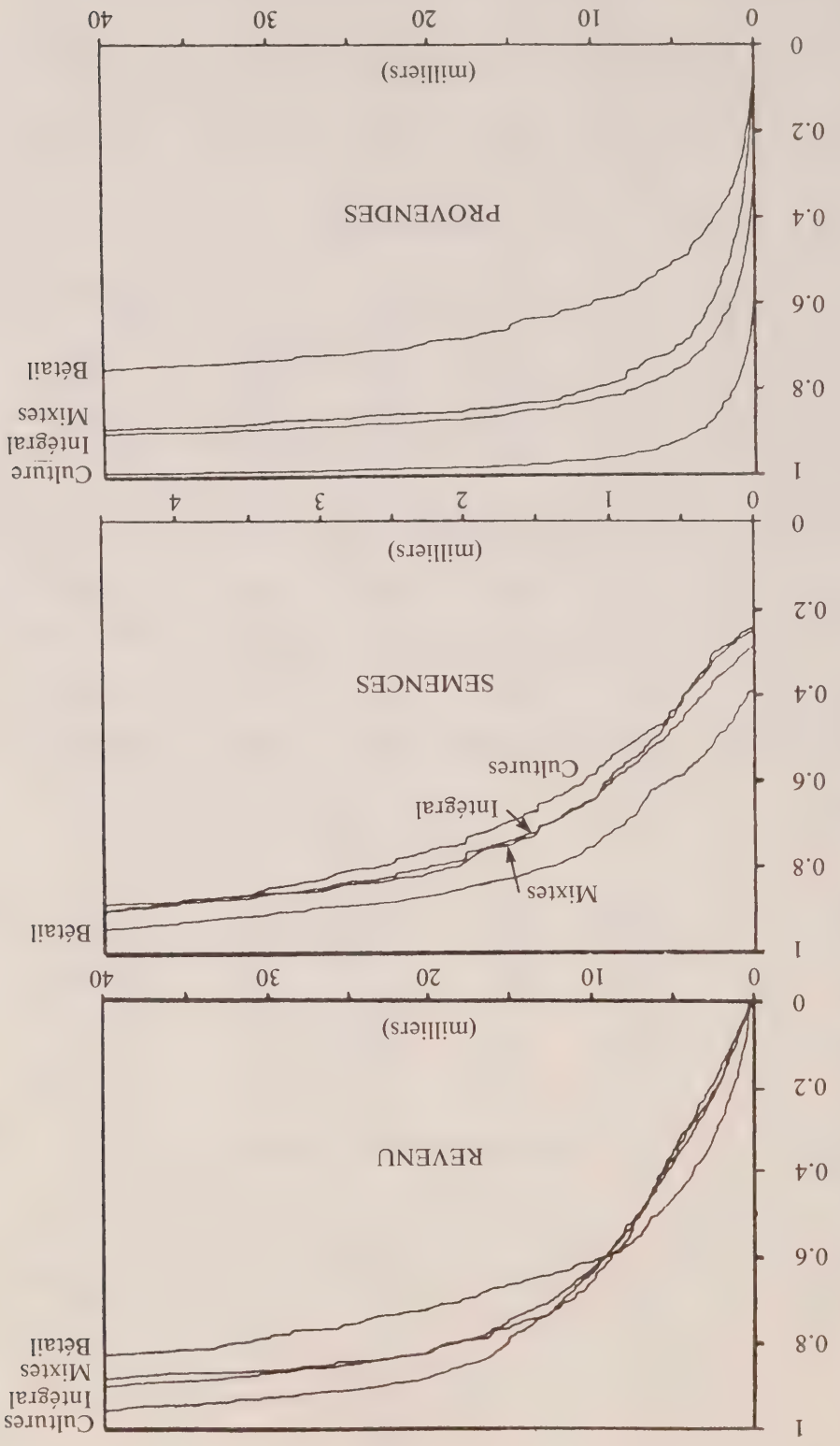


Figure 1: Fonctions de distribution empiriques des variables d'appariement

iiii) Les estimations pondérées ENF pour les totaux des variables pour la province et chaque type de ferme ont été produites à partir de chaque ensemble de données imputées.

iv) Ces étapes ont été répétées dix fois pour obtenir dix répétitions indépendantes (simulations), et on a pris la moyenne des résultats pour chaque variable d'imputation. Ensuite, cette estimation moyenne a été comparée à l'estimation obtenue à partir des fichiers "pro-pres" à la fois au niveau provincial et à celui de chaque type de ferme.

On a recommencé tout le processus avec des taux de non-réponse plus élevés de 15% et de 20% afin d'observer l'impact de ces derniers.

Les variables d'imputation et d'appariement utilisées dans l'étude sont présentées ci-dessous:

Variables d'imputation	
UTIL	= Dépenses en services publics
AUTO	= Dépenses en véhicules agricoles et en machines
TAX	= Impôt foncier
Variables d'appariement	
Type de ferme (appariement exact)	
FEED	= Dépenses en provendes
SEED	= Dépenses en semences
INCOME	= Recettes agricoles brutes

Par ailleurs, le type d'échantillon du donneur était limité par celui du receveur. Il ne faut pas oublier que trois types d'échantillon sont utilisés dans l'ENF: les échantillons spécifiques, les échantillons de listes et les échantillons aréolaires. Il est possible d'imputer une ferme spécifique par une ferme provenant de n'importe quel type d'échantillon, mais ce n'est pas le cas d'un donneur en ce qui concerne une ferme de liste ou aréolaire. De même, une ferme provenant de l'échantillon de liste peut être imputée à partir d'une ferme de l'un des échantillons de liste ou aréolaire, mais ne peut être un donneur que pour les fermes qui font partie de l'échantillon de liste ou qui sont spécifiques. Enfin, les fermes de l'échantillon aréolaire peuvent être imputées uniquement par une autre ferme aréolaire, mais peuvent servir de donneurs pour chacun des trois échantillons. Ces restrictions reposent sur l'hypothèse que si une ferme de liste ou spécifique était utilisée pour une imputation pour une ferme aréolaire, la valeur imputée pourrait relever les estimations d'enquête à un niveau inacceptable en raison des poids d'échantillonnage plus élevés associés aux fermes aréolaires.

3.4 Les fonctions de distribution empiriques de la variable d'appariement

La figure 1 présente les fonctions de distribution empiriques non pondérées des trois variables d'appariement obtenues à partir des groupes d'imputation définis par l'échantillon intégral et selon le type de ferme. À noter que les différences sont appréciables mais pourraient se traduire par la sélection de différents enregistrements donneurs pour un receveur donné.

3.5 Résultats

Les résultats sont présentés au tableau 1. Pour chaque variable d'imputation (UTIL, AUTO ou TAX), chacun des deux ensembles de groupes d'imputation (intégral et par type) et chaque niveau de taux de non-réponse (10%, 15% ou 20%), on a calculé la valeur moyenne des 10 estimations pour le total de la variable pour les dix répétitions. Le biais de cette valeur moyenne est présenté comme un pourcentage de l'estimation "propre". Le cv moyen sur les dix répétitions est également présenté comme un pourcentage.

3. ETUDE EMPIRIQUE

3.1 Justification

En retenant la méthode de l'imputation du plus proche voisin pour l'ENF, il faut régler certains problèmes touchant les détails de la mise en application de cette méthode, en particulier pour ce qui est de la transformation des variables d'appariement. La méthode de la transformation uniforme dans l'imputation N-N pourrait être appliquée en utilisant tous les enregistrements de l'échantillon, ou en n'utilisant que des sous-ensembles de données de l'échantillon. Un groupe de répondants unitaires pour lequel on procède à l'imputation de la non-réponse est appelé un groupe d'imputation.

Des groupes d'imputation différents donneraient des valeurs transformées différentes, les- quelles à leur tour se traduiraient par une sélection différente des enregistrements des donneurs. On s'est posé la question de savoir si la transformation des variables d'appariement au sein d'un groupe d'imputation défini par un critère d'homogénéité qui se rattache étroitement au poste imputé entraînerait une meilleure échelle des variables d'appariement et par- tant, de meilleures données imputées. Ainsi, dans l'ENF, on pourrait s'attendre à ce que la transformation des variables d'appariement au sein des groupes d'imputation définis par le type de ferme donnerait des données imputées meilleures et par conséquent, de meilleures estimations, au sens d'une réduction du biais et de la variance. Malheureusement, la transfor- mation des variables d'appariement coûte cher en termes de ressources informatiques. Si l'on n'a pas besoin de transformer à l'intérieur des groupes d'imputation homogène, il est possi- ble d'économiser sur les coûts informatiques.

Le but principal de cette étude est de répondre à la question suivante dans un cadre ex- périmental: "les deux méthodes de transformation des variables d'appariement, c'est-à-dire la transformation utilisant tous les enregistrements et celles à l'intérieur des groupes de type de fermes, donnent-elles des estimations d'enquêtes sensiblement différentes? Si oui, quelle méthode donne les meilleures estimations?"

3.2 Données utilisées dans l'étude

Après consultation de spécialistes, on a retenu l'échantillon de l'ENF de 1984 pour la pro- vince de l'Alberta. Cet échantillon d'environ 2,000 fermes comprend 50% de fermes de cultures, 27% de fermes d'élevage et 23% de fermes mixtes. Les pourcentages de la popula- tion des trois types de fermes ont été estimés à 52%, 27% et 21% respectivement. Les types de fermes ont été attribués selon la source principale des recettes agricoles projetées de l'ex- ploitation. Si au moins 75% des recettes agricoles projetées d'une ferme provenaient de son stock de bétail, la ferme était classée comme ferme de bétail. Une règle semblable a été utilisée pour la classification des fermes de cultures. Les autres fermes ont été classées comme des fermes mixtes.

3.3 Méthode de l'étude

Nous avons supposé que les données étaient "nettes", même si elles contenaient des valeurs provenant de la procédure d'imputation séquentielle hot-deck. Une fois les données classées selon le type de ferme, on a utilisé la procédure suivante:

i) Dix pour cent des valeurs de chaque variable d'imputation ont reçu au hasard une valeur manquante pour chaque type de ferme. Ces erreurs ont été établies indépendamment pour chaque variable d'imputation.

ii) Les non-réponses obtenues ont été imputées en utilisant la méthode d'imputation N-N sur les deux ensembles de groupes d'imputation définis par l'échantillon intégral (appelé "intégral") et le type de ferme (appelé "type"). Les procédures d'imputation ont été exécutées grâce au système numérique de vérification/imputation (Statistique Canada 1982) dans le progiciel statistique P-STAT (Buhler et Buhler 1978).

estimations de l'enquête serait disproportionnée. Enfin, l'adoption de la nouvelle méthode d'imputation pour l'ENF contribuerait à normaliser la méthodologie d'enquête de toutes les enquêtes agricoles, ce qui est un objectif à long terme de Statistique Canada. À l'heure actuelle, le recensement de l'agriculture et l'enquête sur les données fiscales agricoles utilisent la nouvelle méthode.

La présente communication fait rapport sur une étude empirique qui essaye de fournir des renseignements qui faciliteront la mise en oeuvre de la nouvelle méthode d'imputation. La section suivante décrit brièvement la méthode d'imputation N-N adoptée ici. La section trois présente la procédure d'étude et les principaux résultats obtenus. La section quatre présente nos conclusions provisoires tirées des résultats.

2. METHODE D'IMPUTATION DU PLUS PROCHE VOISIN

La méthode de l'imputation des donneurs consiste, en général, à remplacer les valeurs manquantes ou invalides d'un répondant (receveur) par la réponse valide d'un autre répondant (donneur), dont on suppose qu'il a les mêmes caractéristiques que le receveur. La méthode d'imputation séquentielle hot-deck identifie en séquence les donneurs lors du traitement comme étant ceux qui déclarent les mêmes valeurs que le receveur dans les variables d'appariement spécifiées à l'avance. Cette méthode, cependant, empêche souvent d'obtenir un appariement exact, lorsqu'une variable d'appariement prend un grand nombre de valeurs possibles. Afin de réduire le problème, l'étendue de la variable d'appariement est partagée en intervalles, et l'on obtient le donneur par appariement sur le code d'intervalle. Dans le cas de l'imputation du plus proche voisin, on résout le problème en sélectionnant un donneur à partir d'une mesure de distance multivariée qui représente le degré de similitude entre le donneur et le receveur, défini par les variables d'appariement spécifiées au préalable. Plus deux répondants sont semblables par rapport aux variables d'appariement, et plus la distance est petite. Ainsi, le meilleur donneur pour un receveur est le donneur dont la valeur de la distance par rapport à ce dernier est la plus petite, c'est-à-dire son voisin le plus proche au sens de la distance statistique.

La méthode d'imputation du plus proche voisin utilisée dans la présente étude a été proposée par Sande (1976, 1981). Cette méthode utilise la norme maximum basée sur les données transformées comme la fonction de distance. Voici une brève description de la méthode. Soit $X = (x_1, x_2, x_3, \dots, x_k)$ un vecteur de k variables d'appariement. Chaque variable d'appariement x_j est transformée par $t_j = F(y)$, où $F(y)$ est la fonction de distribution empirique de x_j . Notons que t_j suit la distribution uniforme sur $[0, 1]$. Par conséquent, la distance entre un receveur X^r et un donneur X^d définie par la norme maximum est

$$d(X^r, X^d) = \max |t_j^r - t_j^d|,$$

où t_j^r et t_j^d sont les valeurs transformées de la $j^{ième}$ variable d'appariement x_j dans X^r et X^d , respectivement. Le candidat donneur avec la valeur d la plus petite sera sélectionné, et sa réponse sera inscrite pour la réponse manquante du receveur. La transformation uniforme peut être considérée comme une méthode objective de classement des variables d'appariement quelle que soit leur distribution naturelle.

Étude des effets des groupes d'imputation dans la méthode d'imputation du plus proche voisin pour l'enquête nationale sur les fermes

SIMON CHEUNG et CRAIG SEKO¹

RÉSUMÉ

Un nouveau système de traitement, utilisant la méthode d'imputation du plus proche voisin (N-N), est employé pour l'enquête nationale sur les fermes (ENF). Une étude empirique a été faite pour déterminer si les estimations ENF seraient affectées par l'emploi de groupes d'imputation basés sur le type de ferme. Pour la règle d'imputation examinée ici, l'étude prouve que l'effet peut être petit.

MOTS CLÉS: Enquête nationale sur les fermes; non-réponse; imputation du plus proche voisin; transfor-

mation de la variable d'appartenance.

1. INTRODUCTION

L'enquête nationale sur les fermes (ENF) est une enquête annuelle polyvalente de l'activité agricole au Canada. Cette enquête utilise un plan de sondage à deux bases, c'est-à-dire une base de sondage des grandes fermes (étalée à partir du recensement quinquennal de l'agriculture) et une base de sondage aréolaire des terres agricoles. Les plus grandes unités de la base de sondage de liste sont échantillonnées avec certitude (probabilité de un) en raison de leur impact disproportionné sur les estimations d'enquête. Ces unités sont appelées des fermes spécifiées. Les autres fermes de la base de sondage font l'objet d'une stratification et d'un échantillonnage. Les petites fermes de la population enquêtée, qui sont en comparaison très nombreuses, sont regroupées dans la base de sondage aréolaire et échantillonnées de façon moins complète que les fermes de la base de sondage de liste. Trois échantillons sont donc sélectionnés: un échantillon spécifié, un échantillon de liste et un échantillon aréolaire. Le plan de sondage détaillé de l'ENF a été décrit par Davidson et Ingram (1983) et Davidson (1984). L'ENF fait l'objet d'un traitement par un système adopté à partir des enquêtes précédentes. Ce système utilise la méthode d'imputation séquentielle hot-deck pour corriger la non-réponse des unités et des postes (Phillips 1979). Un nouveau système de traitement sera mis en oeuvre en 1987 afin d'intégrer toutes les enquêtes agricoles effectuées par Statistique Canada. Ce système utilisera la méthode d'imputation du plus proche voisin (N-N) afin de corriger la non-réponse. La décision de mettre en oeuvre cette méthode se justifie par de nombreuses raisons, dont trois sont importantes. D'abord, l'utilisation de la méthode se justifie d'avan-

tage théoriquement que la méthode séquentielle hot-deck avec appariement exact, puisque l'enquête recueille essentiellement des données quantitatives. Ensuite, les études empiriques telles que celles de Kovar (1982) indiquent que les deux méthodes d'imputation donneraient des estimations semblables pour l'ENF et que la méthode N-N se traduirait par un moins grand nombre de points aberrants, c'est-à-dire des données imputées dont la contribution aux

¹ Simon Cheung et Craig Seko, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11^e étage, immeuble R.H. Coats, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

BIBLIOGRAPHIE

- BALLAR, B.A., BAILEY, L., et CORBY, C. (1978). A comparison of some adjustment and weighting procedures for survey data. Dans *Survey Sampling and Measurement* (Ed. N. Krishnan Namboodiri), New York: Academic Press, 175-198.
- CHAPMAN, D.W. (1983). An investigation of nonresponse imputation procedures for the health and nutrition examination survey. Dans *Incomplete Data in Sample Surveys*, Volume 1 - Report and Cases Studies (Ed. W.G. Madow, H. Nisselson, et I. Olkin), New York: Academic Press, 435-483.
- KISH, L. (1978). On the future of survey sampling. Dans *Survey Sampling and Measurement* (Ed. N. Krishnan Namboodiri), New York: Academic Press, 13-21.
- OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys*, Volume 2 - Theory and Bibliographies (Ed. W.G. Madow, I. Olkin, et D.B. Rubin), New York: Academic Press, 143-184.
- PANEL ON INCOMPLETE DATA (1983). Part I - Report. Dans *Incomplete Data in Sample Surveys*. Volume 1 - Report and Cases Studies (Ed. W.G. Madow, H. Nisselson, et I. Olkin), New York: Academic Press, 3-103.
- PLATEK, R., SINGH, M.P., et TREMBLAY, V. (1978). Adjustment for Nonresponse in Surveys. Dans *Survey Sampling and Measurement* (Ed. N. Krishnan Namboodiri), New York: Academic Press, 157-174.
- PLATEK, R., et GRAY, G.B. (1983). Part V - Imputation Methodology: Total Survey Error. Dans *Incomplete Data in Sample Surveys*, Volume 2 - Theory and Bibliographies (Ed. W.G. Madow, I. Olkin, et D.B. Rubin), New York: Academic Press, 249-333.
- TREMBLAY, V. (1975). On the improvement of sample surveys estimates. *Survey Methodology*, 1, 181-196.

Dans la constitution des classes de pondération, il y a donc avantage à tenter de tirer profit d'une part de variables fortement corrélées au phénomène étudié et d'autre part de variables affranchant à la fois une bonne corrélation et un excellent taux de réponse. De plus en croisant entre elles les variables pertinentes on peut former des classes plus homogènes et augmenter conséquemment la valeur du χ^2 . Evidemment le degré de raffinement des classes doit respecter le critère limite exprimé précédemment par l'équation (4.3).

Considérons par exemple, la formation de classes de pondération à partir de trois variables explicatives de l'intention de vote: l'identification du chef de parti qui ferait le meilleur ministre (3 catégories de réponse), le degré de satisfaction face au gouvernement (4 catégories) et la langue maternelle (2 catégories). A ce stade, l'idée est de projeter l'intention de vote mesurée auprès des répondants appartenant à une classe, à l'ensemble des individus de cette même classe, c'est-à-dire ceux dont la classification a pu être établie. La première étape du processus consiste à construire les plus fines classes possibles à partir des 3 variables en cause et à produire un tableau croisé de l'intention de vote selon ces 24 (i.e., $3 \times 4 \times 2$) classes. A l'aide du critère explicite par l'équation ou encore du tableau 4, on élimine par fusionnement les classes de trop petite taille. Lorsque nécessaire, on regroupe donc les classes "semblables", c'est-à-dire celles qui présentent un profil d'intention de vote similaire. On est en mesure alors de produire un tableau comme celui de la page suivante où l'intention de vote apparaît croisée suivant cette nouvelle partition. L'examen des données peut encore suggérer quelques regroupements. De plus le Tableau 6 présente d'autres données pertinentes. Par exemple les deux dernières lignes confrontent par classes le nombre de répondants à la question sur les intentions de vote en regard du nombre total d'individus du sondage classifiées selon les trois variables en cause. On en déduit un premier système de poids. Globalement dans l'exemple il y a 283 personnes dont la classification est connue mais non l'intention de vote. De plus, la valeur globale du χ^2 s'établit à 891 ce qui s'avère une nette amélioration par rapport aux variables prises isolément selon le Tableau 5.

Enfin sous B_j , apparaissent pour chaque P_j à estimer le pourcentage de la variance attribuable à la variance inter-classes. Ces B_j mesurent le gain en précision (réduction de variance) attribuable à l'ajustement les données selon la partition retenue; ceci est évident en réécrivant l'équation (4.1) sous la forme (en négligeant les variances relatives):

$$\text{Var } p_j^* = \text{Var } p_j - (1 - f) B_j \text{Var } p_j$$

Avec un B_j égal à 61,9% pour l'estimation de l'intention de vote à l'endroit du PQ, cela revient à dire que du point de vue de la réduction de la variance l'ajustement des données est équivalente à avoir récupéré sur le terrain 61,9% des 283 non-réponses à la question sur les intentions de vote.

Le problème résiduel est maintenant le suivant: comment ajuster la non-réponse au niveau de questions spécifiques à l'aide de variables elles-mêmes affectées par la non-réponse?

Dans l'exemple mis en lumière par la partition selon le Tableau 6, il est clair qu'une portion significative de la non-réponse sur l'intention de vote ne se trouve pas corrigée par ce genre de pondération. Il reste en effet, 409 cas de non-réponse qui ne peuvent recevoir ce premier traitement puisque pour ceux-ci on ne connaît pas la classification suivant l'une ou l'autre des variables de référence. Une solution à explorer dans ce cas serait d'implanter un système de poids qui permet d'utiliser pour chaque non-répondant le maximum de variables disponible pour estimer la donnée manquante. Par exemple, le profil d'intention de vote de personnes qui n'ont ni répondu à la question portant sur les intentions de vote ni à celle portant sur le meilleur premier ministre mais dont on sait qu'elles sont francophones et qu'elles se sont dites satisfaites du gouvernement au pouvoir serait déduit à partir du profil d'intention de vote des répondants francophones satisfaits du gouvernement. Ce processus d'imputation peut se traduire aisément par un système de poids.

6. APPLICATION ET PROBLÈMES INHÉRENTS

Dans la discussion précédente, nous avons dégagé un critère d'évaluation de la performance de classes de pondération. En pratique, il arrive cependant que les variables qui expliquent le mieux la variance d'un phénomène soient elles aussi entachées du problème de la non-réponse. Cette situation complique alors quelque peu le choix des classes de pondération. Le tableau suivant présente une liste de variables jugées a priori potentiellement intéressantes par un chercheur en vue d'effectuer une pondération d'ajustement pour la non-réponse à la question d'intention de vote. Pour chaque variable candidate, on y décrit la valeur du χ^2 , le nombre de valeurs manquantes et le nombre total de valeurs manquantes lorsqu'elle est croisée avec la question portant sur les intentions de vote. Rappelons que prise isolément cette dernière question cumulait 619 non-réponses dans le sondage.

La valeur du χ^2 est fort révélatrice de la force de prédiction des diverses variables en cause. On observe par exemple que parmi les variables socio-démographiques seule la langue maternelle a un impact qui mérite d'être souligné. Par contre, certaines questions thématiques démontrent un lien non équivoque avec l'intention de vote; en particulier, le degré de satisfaction face au gouvernement actuel et la perception de qui entre les deux principaux chefs de parti ferait le meilleur premier ministre. On se rend bien compte que plus une question est perçue comme parente de la question de fond, plus il s'avère difficile d'obtenir des réponses. Sur le thème plus anodin de la satisfaction face au gouvernement, on n'enregistre que 56 non-réponses (environ 3% de l'échantillon) alors que ce nombre atteint 392 si l'on demande d'identifier le meilleur premier ministre!

Tableau 5
Liste de variables candidates pour compenser, par pondération, l'effet de la non-réponse à la question d'intention de vote

Variable ^a	Valeur du χ^2	Nombre de données manquantes sur la variable	Nombre de données manquantes dans le croisement avec l'intention de vote
Age (6)	34	4	620
Scolarité (4)	8	3	621
Langue maternelle (2)	96	0	619
Degré de satisfaction face au gouvernement du Québec (4)	382	56	625
Degré de satisfaction face au gouvernement du Québec (2)	346	56	625
Identification du meilleur premier ministre (3)	773	392	686
Vote provincial aux élections de 1981	109	269	658
Intérêt face à la politique	1	1	619
Degré de satisfaction à l'endroit du gouvernement fédéral (4)	39	58	631
Intention de vote au fédéral (4)	288	694	832

^a Les nombres entre parenthèses indiquent le nombre de classes considérées pour les variables en cause.

Il est intéressant de rappeler ici que ces résultats ont été développés par analogie dans le contexte de l'échantillonnage en deux phases et que les règles qui ont été déduites peuvent aussi bien s'appliquer aux estimateurs quotients séparés ainsi qu'aux estimateurs poststratifiés. Par exemple, il est souvent utile de savoir jusqu'où le raffinement d'une poststratification produit des résultats plus précis; les règles énoncées ici peuvent donc servir de guide.

5. CRITÈRE DE CHOIX DES VARIABLES D'AJUSTEMENT

Si l'on poursuit l'exemple qui accompagne le présent article, on constate que le degré de satisfaction face au gouvernement peut certainement servir de variable d'ajustement à la non-réponse à la question de l'intention de vote. Mais est-ce vraiment la meilleure variable à utiliser? Si l'instrument d'enquête comporte d'autres questions reliées indirectement à l'intention de vote, à partir de quel critère choisir entre par exemple la satisfaction, certains profils socio-démographiques (langue, scolarité), ou la perception de celui qui ferait le meilleur premier ministre.

Les deux sections précédentes nous enseignent que plus les classes construites sont homogènes plus la variance des estimations ajustées est réduite et plus il est vraisemblable que le biais lui-même s'en trouve amoindri. On a donc avantage à constituer des classes qui maximisent la variance inter-classes de l'estimateur p_j . Si l'on se réfère à l'expression algébrique (4.1), la partition choisie doit maximiser la quantité

$$INTERCL_j = \sum_c (P_{jc} - P_j)^2 P_c$$

Pour une variable X multinationale avec paramètres $P_1, P_2, \dots, P_j, \dots, P_J$, le problème revient à trouver une statistique qui incorpore l'ensemble des quantités $INTERCL_j$ ($j = 1, \dots, J$). Dans ce cas, le χ^2 mérite d'être considéré comme un candidat intéressant puisque

$$\chi^2 = N \sum_j \sum_c (P_{jc} - P_j)^2 P_c / P_j = N \sum_j P_j [INTERCL_j / P_j^2]$$

c'est-à-dire que le χ^2 est égal à une combinaison linéaire des valeurs relatives des $INTERCL_j$ pondérées en fonction des P_j . Par contre, comme $B_j = INTERCL_j / P_j (1 - P_j)$ mesure la proportion de la variance expliquées par la partition en classes, il serait aussi justifié de considérer la statistique

$$\sum B_j = \sum_j \sum_c (P_{jc} - P_j)^2 P_c / P_j (1 - P_j).$$

Notons que cette dernière statistique est équivalente au χ^2 dans trois situations particulières: a) lorsque X est dichotomique b) lorsque les P_j sont à peu près égaux et c) lorsque les P_j sont tous petits. Dans le cas multinominal où il est important de raffiner l'estimation d'un P_j pour un indice j en particulier on peut alors dichotomiser la variable X en fonction de cet indice j et utiliser le comme critère de performance d'une partition en classes. Dans la poursuite de l'exemple nous utiliserons le χ^2 étant donné que cette statistique est produite directement par la plupart des logiciels de traitement de données d'enquêtes.

Tableau 4.
Valeurs de $DIFMIN = \frac{1}{2}\sqrt{a(1 - a)n_c}$ (en %)

n	$a = 1/2$	$a = 1/4$	$a = 1/10$
1000	3.1%	3.7%	5.3%
400	5.0%	5.8%	8.3%
200	7.1%	8.2%	11.8%
100	10.0%	11.5%	16.7%
80	11.2%	12.9%	18.6%
60	12.9%	14.9%	21.5%
40	15.8%	18.3%	26.4%
20	22.4%	25.8%	37.3%
15	25.8%	29.8%	43.0%
10	31.6%	36.5%	52.7%

L'inégalité met donc en lumière une règle simple et suffisante pour qu'il soit avantageux de découper une classe c en $c_1 \cup c_2$. Comme on pouvait s'y attendre intuitivement, plus le nombre de répondants dans la classe c est grand (dans l'échantillon S_1) ou plus les proportions P_{jc_1} et P_{jc_2} s'écartent l'une de l'autre, plus on a intérêt à partitionner finement les classes. Le Tableau 4 au-dessus présente les écarts minimaux $DIFMIN$ correspondant à diverses valeurs de n_c et a .

A titre d'exemple, le tableau précédent nous dit que si l'on a une classe de 100 répondants que l'on considère scinder en deux parties à peu près égales, il faudra que le pourcentage de ceux qui ont le caractère j diffère d'au moins 10% entre les deux nouvelles classes pour que ce raffinement de la partition contribue à réduire l'erreur d'échantillonnage. Si l'écart entre les deux pourcentages est inférieur à 10%, le raffinement est inutile et peu même augmenter la variabilité des estimations produites. On constate par ailleurs, que si les sous-classes c_1 et c_2 sont, très inégales, l'exigence sur le comportement différencié de leurs répondants par rapport au caractère j (i.e., P_{jc_1} versus P_{jc_2}) est plus forte: ainsi, si c_1 représente environ 10% des effectifs de c , l'écart minimal $DIFMIN$ s'établit à 16.7%.

Dans le cas particulier où le découpage d'une classe c s'effectuerait à peu près également entre c_1 et c_2 , l'écart minimal $DIFMIN$ prend une forme très compacte:

$$DIFMIN = 1 / \sqrt{n_c}$$

Dans les situations où le découpage de c se réalise en plusieurs composantes (disons $c = c_1 \cup c_2 \cup \dots \cup c_k$) on peut appliquer le test décrit ici en considérant d'une part la plus petite des sous-classes c_j et d'autre part l'ensemble de toutes les autres. Comme dans ce cas a (ou $1 - a$) risque d'être petit, on peut simplifier la règle énoncée par l'inégalité (4.3) et considérer comme suit l'écart minimal:

$$DIFMIN = \frac{1}{2} \sqrt{\min_j (n_{c_j})}$$

simple permettant d'identifier laquelle des deux partitions C' ou C'' produit la variance la plus faible en tenant compte des facteurs r_c dans l'expression ci-haut. On a que:

$$\begin{aligned} \text{Var } p'' &< \text{Var } p' \\ \Leftrightarrow G &= \sum_{c \in C''} (P_{jc} - P_j)^2 P_c - \sum_{c \in C'} (P_{jc} - P_c)^2 P_c \\ &> \sum_{c \in C''} r_c P_{jc} (1 - P_{jc}) P_c - \sum_{c \in C'} r_c P_{jc} (1 - P_{jc}) P_c = D. \end{aligned}$$

Le membre de gauche G de l'inégalité peut s'exprimer ainsi:

$$G = \sum_{c \in C''} P_j^2 P_c - \sum_{c \in C'} P_j^2 P_c = P_j^2 P_{c_1} + P_j^2 P_{c_2} - P_j^2 P_c \quad (4.2)$$

Si la classe c a été partitionnée ainsi:

$$n_{c_1} = a n_c \text{ et } n_{c_2} = (1 - a) n_c \text{ où } 0 < a < 1$$

alors on a que $P_{jc} = a P_{jc_1} + (1 - a) P_{jc_2}$ que $P_{c_1} = a P_c$ et finalement que $P_{c_2} = (1 - a) P_c$. Alors l'expression (4.2) peut s'écrire sous la forme compacte:

$$G = P_c a (1 - a) [P_{jc_1} - P_{jc_2}]^2.$$

D'autre part, le membre de droite peut se réduire à

$$D = r_{c_1} P_{jc_1} (1 - P_{jc_1}) P_{c_1} + r_{c_2} P_{jc_2} (1 - P_{jc_2}) P_{c_2} - r_c P_{jc} (1 - P_{jc}) P_c.$$

Maintenant, en substituant les variances relatives r_c par l'expression correspondante établie précédemment et remarquant que les termes P_{c_1} , P_{c_2} et P_c sont négligeables par rapport à 1 lorsque l'on s'interroge sur la pertinence du raffinement d'une partition, on obtient:

$$D = ((1/n) [P_{jc_1} (1 - P_{jc_1}) + P_{jc_2} (1 - P_{jc_2}) - P_{jc} (1 - P_{jc})])$$

À cause de la convexité de la fonction $P(1 - P)$ et du fait que P est une combinaison linéaire de P_{jc_1} et P_{jc_2} , la valeur de D est bornée supérieurement par $1/4n$. Ainsi pour que la variance de p'' soit inférieure à celle de p' , il suffit que:

$$P_c a (1 - a) [P_{jc_1} - P_{jc_2}]^2 > 1/4n;$$

en estimant P_c par n_c/n à partir du sous-échantillon S_1 , cette condition prend la forme

suivante:

$$DIF = |P_{jc_1} - P_{jc_2}| > \sqrt{1/2 a (1 - a) n_c} = DIFMIN \quad (4.3)$$

Plus précisément imaginons la situation suivante. Un échantillon aléatoire simple S de taille n' nous donne la distribution d'une variable de classification de la population totale avec $N'_c = (N/n')n'_c$ comme l'estimateur du nombre d'unités de la population appartenant à la classe c . Un sous-échantillon aléatoire simple $S_1 \subset S$ de taille $n = fn'$ ($0 < f < 1$) est choisi dans le but de mesurer la distribution d'une autre variable de classification X . Pour chacune des unités de S_1 on connaît sa classification selon les deux variables décrites ci-haut. On désire estimer la proportion P_j d'unités appartenant à la classe j de la variable X . L'estimateur simple déduit de S_1 est

$$p_j = (1/n) \sum_c n_c P_{jc}$$

Par ailleurs l'estimateur quotient séparé (*post-stratifié*) peut s'exprimer sous la forme:

$$p_j = (1/n') \sum_c n'_c P_{jc}$$

Alors que dans l'expression de p_j toutes les unités de l'échantillon S_1 se voient octroyés un poids égal à 1, dans celle de p_j on observe que les unités ont des poids variables selon qu'elles appartiennent à l'une ou à l'autre des classes c . Ces poids *correctifs* égaux à n'_c/n' utilisent l'information complémentaire disponible auprès de la totalité de l'échantillon S sur la partition en classes.

Si l'on développe la formule de la variance de p_j en conservant les termes de l'ordre de grandeur de la variance relative des N'_c on obtient selon Tremblay (1975):

$$\text{Var } p_j = \text{Var } p_j - [(1-f)/n] \left[\sum_c (P_{jc} - P_j)^2 P_c - \sum_c r_c P_{jc} (1 - P_{jc}) P_c \right] \quad (4.1)$$

où $r_c = N(1 - P_c)/nN'_c$: la variance relative de l'estimateur de N'_c i.e. $N'_c = (N/n')n'_c$

$P_c = N'_c/N$:Proportion de la population appartenant à la classe c ;

$P_{jc} = E p_{jc}$:Proportion des unités qui ont la caractéristique j dans la classe c ;

$P_j = E p_j$: Proportion des unités de la population qui ont la caractéristique j .

L'équation (4.1) nous révèle que la technique d'ajustement par les classes de pondération est d'autant plus efficace que la variance inter-classe est grande et, conséquemment, que la variance intra-classe est petite. Il est facile de vérifier que dans le cas limite où la variance inter-classe est maximale, c'est-à-dire lorsque tous les P_{jc} sont soit 0 ou 1, alors:

$$\text{Var } p_j = P_j(1 - P_j) / n'$$

c'est-à-dire la variance que l'on aurait obtenu si toutes les n' unités avaient répondu. Cette équation (4.1) nous rappelle de plus que dans la mesure où les variances relatives sont négligeables par rapport à 1, on a avantage à partitionner plus finement l'échantillon en un grand nombre de classes; ce faisant, on augmente la variation inter-classes et on diminue par le fait même la variance de p_j .

Mais le raffinement de la partition est justifié par la présence des variances relatives r_c . Regardons-y d'un peu plus près. Supposons qu'une première partition de l'échantillon en un ensemble de classes C' produit l'estimateur p_j tel que défini précédemment. Supposons ensuite une seconde partition plus fine C'' permettant de construire l'estimateur p'_j ; en imaginant que toutes les classes de C' coïncident avec celles de C'' sauf une classe c qui a été scindée en deux parties c_1 et c_2 (i.e., $c = c_1 \cup c_2$), il serait intéressant de trouver un critère

Il est utile de reformuler $B(X_c)$ ainsi:

$$B(X_c) = N\sigma_c^2 d_c(X, X)$$

où σ_c^2 : est la variance de la caractéristique X à l'intérieur de la classe c ;

et $d_c(X, X) = \alpha_c^{-1}(\alpha_c^x - \alpha_c^x)$ est une mesure standardisée de la distance entre la probabilité de réponse moyenne chez ceux qui ont la caractéristique X et ceux qui ne l'ont pas à l'intérieur de la classe c .

Le biais de non-réponse associé à l'estimation p' de P peut donc s'exprimer sous la forme:

$$B(p') = B(N^{-1} \sum^c X_c)$$

$$= N^{-1} \sum^c N\sigma_c^2 d_c(X, X), \tag{3.2}$$

L'expression (3.2) apporte un argument mathématique à la thèse souvent évoquée que l'on a avantage à construire des catégories les plus homogènes possibles par rapport au phénomène étudié en partitionnant l'échantillon en segments dont certains ont tendance à rassembler des unités qui ont la caractéristique X d'une part et les autres d'autre part.

4. IMPACT DE LA PROCÉDURE D'AJUSTEMENT SUR L'ERREUR D'ÉCHANTILLONNAGE

On sait que le problème de la non-réponse a entre autres comme conséquences d'augmenter l'erreur aléatoire d'échantillonnage suite à la perte du nombre d'observations. Il est révélateur d'étudier jusqu'à quel point la technique d'ajustement dont il est question ici compense cette perte de précision. Plusieurs auteurs parmi ceux cités dans l'introduction, ont fait référence au danger potentiel que représente la présence de poids correctifs trop grands ou trop instables étant basés sur un nombre d'observations par classe trop restreint. De leur côté, Plalek et Gray (1983) ont présenté une expression approximative de la composante de la variance attribuable à la non-réponse après l'application de l'ajustement. Bien qu'instructive sur le comportement général de cette composante de la variance échantillonnale, ce développement mathématique ne met pas en lumière le point critique au-delà duquel un raffinement excessif des classes de pondération a un effet adverse sur la précision des données.

En réalité on se trouve devant la situation suivante. Le sondeur a certaines informations fiables auprès d'un échantillon représentatif de la population qu'il étudie (par exemple satisfactif face au gouvernement), cependant les données qui l'intéressent plus spécifiquement dans son enquête (par exemple, intention de vote) ne sont disponibles qu'auprès d'un sous-échantillon et il aimerait utiliser certaines données issues de l'échantillon de base pour améliorer la précision de ses estimateurs. Que l'on parle de non-réponse au niveau des unités échantillonnées ou à celui de questions spécifiques à l'intérieur de la variance des estimateurs, la situation fondamentale est la même. Du point de vue de la variance des estimateurs, la situation correspond à l'application des estimateurs quotients séparés c'est-à-dire à une poststratification selon des catégories définissables à partir des informations disponibles chez l'échantillon de base. Bien sûr que cette analogie est inacceptable pour analyser l'effet biaisant de la non-réponse puisque l'on ne peut soutenir l'hypothèse que le sous-échantillon des répondants est probabilistiquement représentatif de l'échantillon de base. Cependant lorsque l'on étudie la variance des estimateurs, l'approche analogique s'avère aussi utile que défendable.

Pour répondre adéquatement aux questions, il faut poursuivre le développement théorique touchant l'application des classes de pondération.

3. IMPACT DE LA PROCÉDURE D'AJUSTEMENT SUR LE BIAIS DE NON-RÉPONSE

Le défi le plus difficilement surmontable en ce qui concerne la non-réponse est bien de quantifier le degré de réduction du biais de non-réponse suite à l'application de telle ou telle technique. En effet si cette mesure s'avérerait réalisable, il serait possible de mesurer le biais et conséquemment de produire des estimations non biaisées.

Cette dure réalité ne nous empêche cependant pas de tenter de comprendre davantage les mécanismes sous-jacents à la non-réponse afin de concevoir les instruments les plus aptes à réduire son impact sur la qualité des données.

Une façon d'étudier le problème est de le considérer sous l'angle de la théorie des probabilités de réponse selon laquelle on stipule que chaque unité U_i de la population a une probabilité α_i de répondre à l'enquête (ou à une question spécifique posée) si elle est sélectionnée. Même si cette approche doit supposer que les α_i sont non nuls, la théorie permet de déduire des expressions mathématiques du biais de non-réponse selon l'application de telle ou telle méthode en fonction des observations X_i que l'on désire obtenir et des probabilités de réponse α_i . C'est ce qu'ont fait Platek et Gray (1983); dans le cas de l'estimation du sous-total X_c dans la classe de pondération c en ajustant l'estimation échantillonnale par l'inverse du taux de réponse dans la classe c , ils ont établi que le biais résiduel de non-réponse pouvait s'écrire sous la forme:

$$B(X_c) = \alpha_c^{-1} \sum_{N_c}^{N_c} (\alpha_i - \alpha_c) X_i \quad \text{où } \alpha_c = N_c^{-1} \sum_{N_c}^{N_c} \alpha_i \quad (3.1)$$

et où N_c = Taille de la classe c dans la population.

L'expression rappelle que le biais résiduel de non-réponse n'existe après l'application du facteur de correction, que si à l'intérieur de la classe c , il y a une corrélation entre les probabilités de réponse et la caractéristique mesurée.

Il est intéressant par ailleurs d'explicitier l'expression dans le contexte particulier de données de classification, c'est-à-dire lorsque les $X_i = 0$ ou 1. En se référant à la notation introduite à la section précédente, on peut démontrer que le biais résiduel de X_c après l'application du facteur de correction peut s'exprimer à partir (3.1) de sous la forme:

$$B(X_c) = N_c P_c \alpha_c^{-1} (\alpha_x - \alpha_c) = N_c P_c (1 - P_c) \alpha_c^{-1} (\alpha_x - \alpha_c^2);$$

où P_c = proportion réelle des unités de la classe c qui ont la caractéristique X_i

α_x = moyenne des probabilités de réponse chez les unités de la classe c qui ont la caractéristique X_i

et α_c^2 = moyenne des probabilités de réponse chez les unités de la classe c qui n'ont pas la caractéristique X_i .

Tableau 3

Satisfaction face au gouvernement croisée
selon le fait de répondre ou non
à la question sur les intentions de vote
(nombre de cas pondérés)

Chez les satisfaits		Chez les insatisfaits		TOTAL
Réponse à l'intention de vote		Non-réponse à l'intention de vote		TOTAL
$n_1 = 555$		$n_2 = 656$		$n = 1211$
236		334		570
$n'_1 = 791$		$n'_2 = 989$		$n' = 1780^a$

^a Ce tableau exclut 56 cas de non-réponse à la question portant sur la satisfaction.

Alors que 70,1% des personnes satisfaites du gouvernement avaient l'intention d'appuyer le parti au pouvoir (PQ) la situation est inversée chez les insatisfaits, comme on pouvait s'y attendre, alors que 76,1% d'entre eux prévoyaient voter pour le parti d'opposition d'alors (PLQ).

Pour tirer profit de cette information auxiliaire, l'une des techniques disponibles est de créer des classes de pondération basées sur la satisfaction. Les informations du Tableau 3 présentent les données complémentaires nécessaires pour effectuer les ajustements. Si l'on considère les 'satisfaits' et les 'insatisfaits' comme deux classes de pondération, l'ajustement statistique des données prend la forme suivante:

soient p_{jc} = la proportion des répondants de la classe c qui ont l'intention d'appuyer le parti j ;
 n_c = le nombre de cas dans la classe c qui ont répondu à la question sur les intentions de vote;
 $n = \sum_c n_c$ = la taille du sous-échantillon S_1 des répondants aux questions d'intention de vote et de satisfaction;
 n'_c = le nombre total de cas dans la classe c ;
et $n' = \sum_c n'_c$ = la taille de l'échantillon S des répondants à la question de satisfaction.

Alors, les estimations ajustées de l'intention de vote se calculent ainsi

$$p'_j = \sum_c n'_c p_{jc} (1/n)$$

Cette nouvelle estimation correspond à introduire un poids correctif égal à $n'_c/n \cdot n'$ pour tous les répondants de la classe c .

Ce simple exercice illustre le fonctionnement du mécanisme bien connu de l'ajustement statistique par la construction de classes de pondération inspirée des procédures traditionnelles de poststratification. Les questions qu'il faut approfondir dans une telle application sont les suivantes:

1. Quel est l'impact de cette procédure sur la réduction du biais de non-réponse?
2. Comment cette technique affecte-t-elle l'erreur d'échantillonnage?
3. Quelles sont les variables auxiliaires (ou combinaisons de variables) qui peuvent le mieux servir à la définition des classes?
4. Jusqu'à quel point trouve-t-on avantage à raffiner la définition des classes de pondération?
5. Que faire avec des variables auxiliaires qui comportent elles aussi de la non-réponse?

2. ILLUSTRATION DE LA TECHNIQUE

Imaginons l'exemple bien réel et fréquemment rencontré qu'est celui de la mesure des intentions de vote. Toutes les données utilisées dans ce texte proviennent du sondage OMNIBUS du Centre de sondage de l'Université de Montréal, Automne 1985 dont une section visait à mesurer les intentions de vote aux élections québécoises de décembre 1985, quatre semaines avant l'événement. La répartition des réponses à la question portant sur les intentions de vote parmi les 1 836 personnes interrogées qui avaient l'intention d'aller voter se lisait ainsi: Ce tableau présente une situation évidente où le problème de la réponse ne peut être ignoré. Répartir les non-réponses aveuglément au prorata des autres réponses est un exercice hasardeux qui suppose que ceux qui n'ont pas exprimé leur intention de vote ont le même profil que ceux qui ont répondu spontanément à la question.

Les deux conséquences de la présence aussi marquée de la non-réponse sont bien connues: biais potentiel et augmentation de l'erreur d'échantillonnage suite à la réduction effective de la taille de l'échantillon. Toute technique d'ajustement doit tenter de réduire ces deux effets. Lorsque, comme dans ce cas-ci, l'importance du phénomène de la non-réponse est prévisible, il y a lieu de concevoir le questionnaire de façon à y incorporer des questions corrigées qui peuvent servir de base à des ajustements éventuels. Par exemple, une question permettant de connaître la satisfaction des personnes interrogées face au gouvernement en place peut être d'un secours apprécié étant donnée l'étroite relation existant entre cet indice et l'intention de vote comme en fait foi le tableau suivant.

Tableau 1
Répartition des intentions de vote
(avec non-réponse)

	<i>n</i> ^a	%
Parti Québécois (PQ)	505	27.5
Parti Libéral du Québec (PLQ)	650	35.4
Autres partis	62	3.4
Non-réponse	619	33.7
TOTAL	1 836	100.0

^a Nombre de cas pondérés.

Tableau 2
Intention de vote au provincial selon la
satisfaction face au gouvernement

Intentions de vote	Chez les satisfaits (<i>n</i> = 555)	Chez les insatisfaits (<i>n</i> = 656)
PQ	70.1%	17.3%
PLQ	26.7%	76.1%
Autres	3.2%	6.6%

Critères pratiques pour la définition des classes de pondération

VICTOR TREMBLAY¹

RÉSUMÉ

Lorsque la technique d'ajustement par classes de pondération est mise en application pour compenser l'effet de la non-réponse, plusieurs questions méritent des réponses précises et quantifiées: Comment le choix des variables servant à la définition des classes affecte-t-il l'erreur quadratique moyenne totale, en particulier le biais de non-réponse et la variance échantillonnale? Selon quelle règle et quelle procédure doit-on choisir les variables d'ajustement? À partir de quel critère peut-on établir des tailles optimales pour les classes de pondération? Enfin, lorsque cette procédure est mise en application pour compenser la non-réponse au niveau d'éléments spécifiques d'un questionnaire comment utiliser efficacement des variables auxiliaires fortement corrélées lorsque celles-ci sont elles-mêmes affectées par la non-réponse? Cet article s'adresse aux professionnels impliqués dans la pratique qui cherchent des lignes directrices.

MOTS CLÉS: Ajustement pour la non-réponse; classes de pondération; poststratification; biais de non-réponse.

1. INTRODUCTION

Le problème de l'application de la technique d'ajustement de la non-réponse par la création de classes de pondération a une parenté évidente avec celui de la détermination des critères de poststratification. Kish (1978) a identifié ce domaine de recherche comme pressant en rap- pelant que les effets de ce genre de pondération ont souvent une résultante inconnue lorsque l'on met en balance les avantages et les inconvénients. Au même moment Platek, Singh et Tremblay (1978) ont développé des expressions mathématiques du biais et de la variance des estimateurs découlant de l'ajustement par classes de pondération. Leur modèle basé sur le concept de "probabilité de réponse", a été exploité plus à fond récemment par Platek et Gray (1983). Parallèlement, Bailar, Bailey et Corby (1978) ont décrit la recherche théorique et em- pirique entreprise au US Bureau of the Census et leur présentation se termine en soulignant l'importance et la nécessité de développer des assises théoriques solides touchant les méthodes d'ajustement pour la non-réponse. Plus récemment le "Panel on Incomplete Data" (1983) a brossé un tableau particulièrement concis et complet des implications pratiques de l'ajuste- ment par pondération et a mis en relief les conclusions auxquelles sont arrivés Oh et Scheuren (1983) suite à une étude de simulation. De son côté, Chapman (1983) a analysé un certain nombre de procédures permettant d'identifier les variables les plus pertinentes pour construire efficacement les classes de pondération.

Le présent article s'inscrit dans la ligne de continuité de ces recherches en tentant de cir- conscrire certaines règles d'application de cette procédure d'ajustement à partir de fondements théoriques. L'unique exemple qui tout au long du texte sert à illustrer la discussion pêche cer- tainement par sa spécificité; mais le lecteur saura certes en dégager des champs d'application beaucoup plus variés et riches.

¹ Victor Tremblay, Président, STATPLUS experts-consultants en statistique, C.P. 337 Ville Mont-Royal, Québec H3P 3C6.

4. CONCLUSION

Pour les enquêtes-entrevues périodiques, il est essentiel d'avoir des systèmes informati-ques qui peuvent contrôler avec rapidité et précision la qualité des données provenant de ces enquêtes. Inversement, lorsqu'il est difficile d'obtenir les données voulues, le système doit pouvoir effectuer le mieux possible une imputation pour la non-réponse suivant des règles bien définies.

Le processus de contrôle permettra de déceler des enregistrements erronés. Les erreurs relevées peuvent être graves ou non. Il serait souhaitable de corriger toutes les erreurs soit en revoquant les questionnaires ou en vérifiant l'authenticité des réponses auprès des répondants. Si ces opérations ne peuvent être effectuées à cause des délais fixés ou de contraintes budgétaires, il faut au moins corriger les erreurs les plus graves. Ensuite, il faut procéder à l'imputation pour la non-réponse. Un résumé des interventions (contrôle ou imputation) du système devrait être produit afin de renseigner l'analyste d'enquête sur l'état des données.

BIBLIOGRAPHIE

- BERTHELLOT, J.-M., et HIDIROGLOU, M.A. (1982). Specifications for imputations in the retail trade survey. Document technique de Statistique Canada.
- BERTHELLOT, J.-M. (1983). Méthodes de vérification statistique pour l'enquête sur le commerce de gros et le commerce de détail. Document technique de Statistique Canada.
- DIXON, W.G. (1953). *Processing data for outliers*. dans Biometrics, vol. 9, n° 1, p. 74-89.
- ERNST, L.R. (1980). Comparison of Estimators of the Mean which Adjust for Large Observations. *Sankhya*, 42, p. 1-16.
- FELLEGLI, I.P., et HOLT, D. (1976). *A systematic approach to automatic edit and imputation* dans Journal of the American Statistical Association, 71, p. 17-35.
- GENTLEMEN, J.F., et WILK, M.B. (1975). *Detecting outliers, II. Supplementing the direct analysis of residuals* dans Biometrics, 31, p. 387-410.
- GRUBBS, F.E. (1969). *Procedures for detecting outlying observations in samples* dans Technometrics, 11, p. 1-21.
- GUMBEL, E.J. (1960). *Discussion on "Rejection of outliers" by Anscombe, F.J.* dans Technometrics, 2, p. 165-166.
- HIDIROGLOU, M.A. et SRINATH, K.P. (1981). *Some estimators of population totals from a simple random sample containing large units* dans Journal of the American Statistical Association, 76, p. 690-695.
- KENDALL, M.G., et BUCKLAND, W.R. (1957). A Dictionary of Statistical Terms. New York, Hafner.
- PRESCOTT, P. (1978). *Examination of the behaviour of tests for outliers when more than one outlier is present* dans Applied Statistics, vol. 27, n° 1, p. 10-25.
- SUGAVANAM, R. (1983). A statistical edit for change. Document technique de Statistique Canada.
- SANDE, I.G. (1981). Estimation in the revised ISPL. Document technique de Statistique Canada.
- TIETGEN, G.L., et MOORE, R.H. (1972). *Some Grubbs - type statistics for the detection of several outliers* dans Technometrics, 55, p. 583-598.
- WILKINSON, R.G. (1982). An outlier identification technique designed for the Business Finance Annual Survey. Document de Statistique Canada.

$$(c) \quad s_{i,NR}(t) = s_{i,NR}(t-1) [x_{ip}(t)/x_{ip}(t-1)]$$

$$(d) \quad s_{i,NR}(t) = x_{ip}(t) - s_{i,R}(t).$$

Parmi ces formules, nous choisirons la première, dans l'ordre précité, qui satisfera à la condition d'inégalité.

Pour toutes les situations ci-dessus, les valeurs imputées (réelles) seront donc calculées comme suit:

$$I_{(2)}^{ij}(t) = (1 - \delta_{ij}) [s_{i,NR}(t)/s_{i,NR}(t-1)] x_{ij}(t-1) + \delta_{ij} x_{ij}(t); j = 1, \dots, p-1$$

Deuxième cas: $x_{ip}(t)$ n'existe pas et il manque des éléments du vecteur élémentaire.

Comme dans le premier cas, deux situations sont possibles:

$$(i) \quad x_{ip}(t) = \sum_{d=1}^f x_{ij}(t)$$

Si $\Sigma_{d=1}^{f-1} \delta_{ij} = 0$, alors $s_{i,NR}(t) = I_{(1)}^{ip}(t)$ où $I_{(1)}^{ip}(t)$ a été obtenue par l'imputation pour la non-réponse totale. Prenons maintenant la formule d'imputation $I_{(2)}^{ij}(t)$. Si $\Sigma_{d=1}^{f-1} \delta_{ij} > 0$, $I_{(2)}^{ij}(t)$ est utilisée à la condition que $s_{i,NR}(t) = I_{(1)}^{ip}(t) - s_{i,R}(t) \geq 0$. Autrement, il faut utiliser la formule d'imputation suivante:

$$I_{(3)}^{ij}(t) = (1 - \delta_{ij}) [s_{i,NR}(t)/s_{i,NR}(t-1)] x_{ij}(t-1) + \delta_{ij} x_{ij}(t); j = 1, \dots, p-1$$

et $I_{(1)}^{ip}(t)$ est remplacée par $I_{(3)}^{ip}(t) = \Sigma_{d=1}^{f-1} I_{(3)}^{dp}(t)$

$$(ii) \quad x_{ip}(t) > \Sigma_{d=1}^{f-1} x_{ij}(t)$$

Pour cette situation, on remplace la variable $x_{ip}(t)$ définie au cas I(ii) par $I_{(1)}^{ip}(t)$ et on applique les méthodes définies pour ce cas à la condition que l'inégalité ci-dessus soit respectée. Si cela n'est pas possible, on doit utiliser $I_{(3)}^{ip}(t)$ et $I_{(1)}^{ip}(t)$ est alors remplacée par $I_{(3)}^{ip}(t) = \Sigma_{d=1}^{f-1} I_{(3)}^{dp}(t)$.

Si l'hypothèse selon laquelle les éléments de données des vecteurs $\tilde{x}_i(t)$ et $\tilde{x}_i(t-1)$ sont répartis de la même façon n'est pas valide, chaque élément doit alors faire l'objet d'une imputation propre selon les méthodes prévues pour la non-réponse totale. Les valeurs imputées doivent ensuite être rajustées pour satisfaire à la condition d'inégalité $x_{ip} \geq \Sigma_{d=1}^{f-1} x_{ij}$. Ainsi, dans le cas I(i), par exemple, nous aurions pour $\Sigma_{d=1}^{f-1} \delta_{ij} = 0$

$$I_{(4)}^{ij}(t) / [x_{ip}(t) - \Sigma_{d=1}^{f-1} I_{(4)}^{dp}(t)] = \sum_{d=1}^f I_{(4)}^{ij}(t) / [I_{(4)}^{ij}(t) + \delta_{ij} x_{ij}(t)]$$

et pour $\Sigma_{d=1}^{f-1} \delta_{ij} > 0$

$$I_{(4)}^{ij}(t) = (1 - \delta_{ij}) \left[\frac{x_{ip}(t) - \Sigma_{d=1}^{f-1} \delta_{ij} x_{ij}(t)}{\Sigma_{d=1}^{f-1} (1 - \delta_{ij}) I_{(4)}^{ij}(t)} \right] + \delta_{ij} x_{ij}(t); j = 1, \dots, p-1.$$

De même, les valeurs imputées $I_{(4)}^{ij}(t)$ pourraient servir à développer les cas I(ii) et 2.

3.1 Non-réponse partielle

Pour un vecteur élémentaire $(x_{i1}(t), x_{i2}(t), \dots, x_{id}(t))$ qui appartient à $\tilde{x}_i(t)$, posons δ_{ij} la variable auxiliaire qui prend la valeur 1 ou 0 selon que $x_{ij}(t)$ existe ou non à la période t . Pour faciliter l'exposé, nous allons définir quelques nouveaux paramètres. À cette fin, posons:

$$s_{i,R}(t-1) = \sum_{j=1}^d \delta_{ij} x_{ij}(t-1)$$

= le nombre total de zones qui contenaient une réponse à la période $t-1$ et qui en contenaient également une à la période t ,

$$s_{i,NR}(t-1) = \sum_{j=1}^d (1 - \delta_{ij}) x_{ij}(t-1)$$

= le nombre total de zones qui contenaient une réponse à la période $t-1$, mais qui n'en contenaient plus à la période t ,

$$s_{i,R}(t) = \sum_{j=1}^d \delta_{ij} x_{ij}(t).$$

L'imputation pour la non-réponse partielle est fondée sur l'hypothèse selon laquelle $x_{ip}(t) \geq \sum_{j=1}^{d-1} x_{ij}(t)$ et celle selon laquelle les éléments de $\tilde{x}_i(t)$ sont répartis de la même façon que les éléments de $\tilde{x}_j(t-1)$ à l'intérieur de leurs vecteurs respectifs. Nous analyserons deux cas particuliers.

Premier cas: $x_{ip}(t)$ existe mais il manque des éléments du vecteur élémentaire.

Deux situations sont possibles: $x_{ip}(t) = \sum_{j=1}^{d-1} x_{ij}(t)$ ou $x_{ip}(t) > \sum_{j=1}^{d-1} x_{ij}(t)$.

$$(i) \quad x_{ip}(t) = \sum_{j=1}^{d-1} x_{ij}(t)$$

S'il manque tous les éléments de $\tilde{x}_i(t)$, à l'exception de $x_{ip}(t)$, c'est-à-dire que $\sum_{j=1}^{d-1} \delta_{ij} > 0$, alors si, $\delta_{ij} = 0$, l'équation suivante doit donc être satisfaite: $s_{i,NR}(t) = x_{ip}(t)$. S'il manque certains éléments de $\tilde{x}_i(t)$, à l'exception de $x_{ip}(t)$, c'est-à-dire que $\sum_{j=1}^{d-1} \delta_{ij} > 0$, alors si,

$$(ii) \quad \sum_{j=1}^{d-1} x_{ij}(t) > x_{ip}(t)$$

S'il manque tous les éléments de $\tilde{x}_i(t)$, à l'exception de $x_{ip}(t)$, alors $s_{i,NR}(t) = s_{i,NR}(t-1) x_{ip}(t) / x_{ip}(t-1)$. S'il manque certains éléments de $\tilde{x}_i(t)$, à l'exception de $x_{ip}(t)$, il n'est pas aussi facile de définir $s_{i,NR}(t)$. Quoi qu'il en soit, l'inéquation suivante doit être satisfaite: $s_{i,R}(t) + s_{i,NR}(t) > x_{ip}(t)$. À cette fin, nous définirons par ordre de préférence quatre formules d'imputation distinctes pour $s_{i,NR}(t)$.

(a) $s_{i,NR}(t) = [s_{i,NR}(t-1) + s_{i,R}(t-1)] x_{ip}(t) / x_{ip}(t-1) - s_{i,R}(t)$ à la condition que $s_{i,NR}(t) \geq 0$. Il est à noter que l'inéquation $x_{ip}(t) > \sum_{j=1}^{d-1} x_{ij}(t)$ est satisfaite si $s_{i,NR}(t) \geq 0$.

$$(b) \quad s_{i,NR}(t) = s_{i,NR}(t-1) [s_{i,R}(t) / s_{i,R}(t-1)]$$

$$\begin{aligned}
 z_{(3)}^{(ip)}(t) &= [1 \sum_{res3} w_r x_{rp}^{ip}(t) / \sum_{res3} w_r x_{rp}^{ip}(t-1)], \\
 z_{(4)}^{(ip)}(t) &= [1 \sum_{res4} w_r x_{rp}^{ip}(t) / \sum_{res4} w_r x_{rp}^{ip}(t-1) x_{ip}^{ip}(t-1)], \\
 z_{(5)}^{(ip)}(t) &= [1 \sum_{res5} w_r x_{rp}^{ip}(t) / \sum_{res5} w_r], \\
 z_{(6)}^{(ip)}(t) &= [1 \sum_{res6} w_r x_{rp}^{ip}(t) / \sum_{res6} w_r],
 \end{aligned}$$

où w_r = l'inverse de la probabilité de sélection de l'unité r pour la cellule donnée. Les sous-ensembles s_l ($l=1, \dots, 6$), sont formés des unités qui ont fourni une réponse pour la *pieme* variable à la période t et qui ont satisfait aux critères de contrôle. Les conditions de formation de chaque sous-ensemble sont les suivantes:

s_1 = toutes les unités qui ont fourni des réponses vérifiées entre les périodes t et $t-1$.

s_2 = toutes les unités qui ont fourni des réponses vérifiées entre les périodes t et $t-\bar{Q}$.

s_3 = les unités du sous-échantillon de suivi qui ont fourni des réponses vérifiées entre les périodes t et $t-1$.

s_4 = les unités du sous-échantillon de suivi qui ont fourni des réponses vérifiées entre les périodes t et $t-\bar{Q}$.

s_5 = toutes les unités qui ont fourni des réponses vérifiées à la période t .

s_6 = les unités du sous-échantillon de suivi qui ont fourni des réponses vérifiées à la période t .

Les facteurs qui déterminent le choix de la méthode d'imputation sont les suivants:

(i) On utilisera la méthode 1 (ou 2) si une réponse a été fournie ou une valeur imputée à la période $t-1$ (ou $t-\bar{Q}$) et si l'on croit que la tendance des données relatives aux non-répondants est la même que celle des données relatives aux répondants à l'intérieur de la cellule donnée.

(ii) On utilisera la méthode 3 (ou 4) si une réponse a été fournie ou une valeur imputée à la période $t-1$ (ou $t-\bar{Q}$) et si l'on croit que la tendance des données relatives aux non-répondants diffère de celle des données relatives aux répondants à l'intérieur de la cellule donnée.

(iii) On utilisera la méthode 5 si aucune réponse n'a été fournie aux périodes $t-1$ et $t-\bar{Q}$ et si l'on croit que la moyenne des données relatives aux non-répondants est égale à la moyenne des données fournies par les répondants à l'intérieur de la cellule donnée.

(iv) Enfin, on utilisera la méthode 6 si aucune réponse n'a été fournie aux périodes $t-1$ et $t-\bar{Q}$ et si l'on croit que la moyenne des données fournies par les répondants diffère de la moyenne des données relatives aux non-répondants.

Le choix de la méthode appropriée peut se faire à l'aide de tables de décision qui définissent les conditions et permettent de choisir, en fonction de ces conditions, la meilleure méthode d'imputation selon des règles prédétablies. Une fois qu'un vecteur élémentaire de $x_{ip}^{ip}(t)$ a fait l'objet d'une imputation, on peut appliquer aux autres vecteurs élémentaires les méthodes d'imputation pour la non-réponse partielle.

• l'ordre d'imputation pour les unités non répondantes doit être le suivant: tendances (mensuelles, trimestrielles, annuelles) et moyennes (médianes), la priorité allant aux tendances les plus récentes. Dans un système d'imputation mensuelle, par exemple, on applique des tendances mensuelles aux unités seulement s'il existe des données (fournies ou imputées) pour le mois précédant celui qui fait l'objet de l'imputation. Les tendances annuelles sont appliquées surtout aux unités qui exercent une activité saisonnière et qui omettent de fournir les renseignements demandés à l'issue de la morte-saison, auquel cas il doit exister des données pour le mois correspondant de l'année précédente. Pour faire l'imputation fondée sur les tendances, on multiplie celles-ci par les données du mois précédent ou de l'année précédente. Dans l'éventualité où on ne pourrait appliquer des tendances, la moyenne (médiane) de la cellule servirait à l'imputation.

Pour exprimer ce qui précède sous forme de modèle mathématique, posons n le nombre d'unités qui sont censées répondre à l'enquête pour une cellule et un mois donnés. Soit n_3 le nombre d'unités qui ne répondent pas du tout au questionnaire, n_1 le nombre d'unités qui répondent entièrement au questionnaire et n_2 le nombre d'unités qui répondent partiellement au questionnaire. Nous supposons qu'il s'agit d'un échantillon stratifié prélevé selon un échantillonnage aléatoire simple sans remise. Soit $m_3 (2 \leq m_3 \leq n_3, m_3$ ayant été prélevé dans n_3 suivant une méthode de répartition au hasard) la taille de l'échantillon de non-répondants qui doivent faire l'objet d'un suivi. Soulignons que n_4 unités ($n_4 = n - \sum_{i=1}^3 n_i$) ne devraient pas être en mesure de participer à l'enquête pour diverses raisons. À la période t , elles pourraient se trouver dans la morte-saison, être inactives ou disparues ou encore ne pas faire partie du champ de l'enquête. Pour ces unités, le système imputera automatiquement des valeurs nulles dans toutes les zones appropriées pour la période donnée. On appliquera ensuite diverses méthodes d'imputation selon le genre de non-réponse.

3.0 Non-réponse totale

Nous analyserons tout d'abord le processus d'imputation pour les unités de non-réponse totale. Comme tout le vecteur $\tilde{x}_i(t)$ ou quelques-uns de ses vecteurs élémentaires définis à la section 2.0 doivent faire l'objet d'une imputation globale, désignons $(x_{i1}(t), \dots, x_{id}(t))$ l'un des vecteurs élémentaires de $\tilde{x}_i(t)$ où le contrôle et l'imputation des données ne dépendent pas des autres vecteurs élémentaires contenus dans $\tilde{x}_i(t)$. En supposant que

$$x_{id}(t) \geq \sum_{j=1}^f x_{ij}(t),$$

(ce qui implique que la somme des $p - 1$ premiers éléments d'information des vecteurs élémentaires est inférieure au *pieme* élément d'information, c'est-à-dire au total)

la première valeur imputée de $x_{id}(t)$ est:

$$I_{(1)}^{(id)}(t) = \sum_{k=1}^6 [Z_{(k)}^{(id)}(t) \delta_{(k)}^i]$$

où $\delta_{(k)}^i$ désigne la méthode d'imputation utilisée et $Z_{(k)}^{(id)}$ est la valeur imputée correspondante. Une des six valeurs de $\delta_{(k)}^i$ est égale à 1 et les cinq autres sont nulles ($\sum_{k=1}^6 \delta_{(k)}^i = 1$). Les valeurs imputées $Z_{(k)}^{(id)}(t)$ sont les suivantes:

$$Z_{(1)}^{(id)}(t) = [\sum_{r \in S_1} w_r x_{rd}(t) / \sum_{r \in S_1} w_r x_{rd}(t-1)] x_{id}(t-1),$$
$$Z_{(2)}^{(id)}(t) = [\sum_{r \in S_2} w_r x_{rd}(t) / \sum_{r \in S_2} w_r x_{rd}(t-1)] x_{id}(t-1),$$

3. IMPUTATION DE DONNÉES PÉRIODIQUES

Les données recueillies périodiquement auprès des entreprises (par exemple, données sur les ventes ou l'emploi) proviennent d'enquêtes par sondage effectuées au moyen d'interviews téléphoniques ou de questionnaires envoyés par la poste. Les unités non répondantes font l'objet d'un suivi le plus rigoureux possible selon les crédits accordés en vue d'accroître les taux de réponse. Le suivi est normalement effectué par courrier dans le cas des petites et moyennes entreprises et par téléphone dans le cas des entreprises de grande taille. Même si le suivi effectué auprès des entreprises retarde parfois les taux de réponse pour une période de référence donnée, il subsistera toujours un groupe d'entreprises non répondantes que l'on pourra classer soit comme entreprises problèmes soit comme entreprises négligentes. Les entreprises problèmes sont des unités qui exigent un grand effort de persuasion avant d'accepter de se prêter à l'enquête, si jamais elles acceptent. Les entreprises négligentes sont des unités qui répondent tardivement à l'enquête par rapport à la période de référence soit parce qu'elles ne retournent pas leur questionnaire à la date fixée soit parce qu'il faut leur rappeler leur obligation au moyen d'un questionnaire de rappel. Ainsi, lorsqu'il y a des unités non répondantes, on doit recourir à l'imputation pour que ces unités ne soient pas exclues du calcul de l'estimateur visé par l'enquête. Dans les enquêtes-entreprises mensuelles, comme l'enquête mensuelle sur le commerce de détail, ce sont des totaux (par exemple, ventes) que l'on estime. L'imputation peut également servir à produire des valeurs pour des unités qui sont désignées comme observations aberrantes. Les valeurs imputées peuvent remplacer les observations aberrantes lorsqu'il n'existe aucune raison valable pour expliquer l'existence de telles observations.

Les unités qui n'auront finalement pas répondu à l'enquête sont appelées des unités de non-réponse totale et celles qui n'auront pas fourni tous les renseignements demandés sont appelées des unités de non-réponse partielle. Un système d'imputation devrait idéalement avoir les propriétés suivantes (Berthelot et Hidiroglou, 1982):

- il doit permettre de déterminer automatiquement la méthode d'imputation la plus appropriée dans une situation donnée;
- la cellule d'imputation, c'est-à-dire le niveau auquel s'effectue le calcul des tendances et des moyennes (médianes), doit normalement correspondre au dernier niveau de stratification de l'échantillon;
- un nombre minimum d'unités doit être inclus dans le calcul des tendances ou des moyennes (médianes), sinon les cellules d'imputation sont combinées automatiquement (suivant un mode prédéterminé) jusqu'à ce qu'on obtienne le minimum d'unités requis;
- le système doit désigner par des codes de situation les unités qui ne doivent pas faire l'objet d'une imputation; parmi celles-ci, il y a les unités à activité saisonnière, durant la morte-saison, les unités qui ne sont pas en exploitation pour une période déterminée ou les unités qui ne sont plus du tout en exploitation;
- les nouvelles entreprises qui n'ont aucune expérience commerciale doivent pouvoir faire l'objet d'une imputation fondée sur les moyennes (médianes) calculées à partir des données fournies par d'autres entreprises récemment créées qui appartiennent à la même catégorie; des unités doivent pouvoir faire l'objet d'une imputation pour un certain nombre de périodes précédant la période courante si des données réelles ont servi à mettre à jour les dossiers de ces unités au cours des périodes en question: cette opération vise à améliorer la valeur des imputations de l'année courante;
- une imputation rétrospective doit pouvoir être appliquée aux unités qui ont toujours fait l'objet d'une imputation prospective, dès qu'on obtient une bonne réponse pour une période donnée;
- les unités qui ont fait l'objet d'une imputation doivent pouvoir être identifiées par une code d'imputation correspondant à la méthode utilisées;

Si l'on veut tenir compte de la valeur des données, il est nécessaire de définir la transformation suivante (Berthelot, 1983):

$$E_i = s_i \{ \text{Max} (x_i(t), x_i(t + 1)) \}^U$$

où $0 \leq U \leq 1$. Les valeurs de E_i représentent des effets et l'exposant U détermine l'importance de la valeur des données. Cette transformation nous permet de prêter plus d'importance à une petite variation touchant une grande unité qu'à une forte variation touchant une petite unité. Les valeurs de la médiane et des quartiles utilisées par Sande (1981) sont appliquées aux valeurs de E_i afin de déceler des observations aberrantes. En désignant respectivement par E_{Q1} , E_M et E_{Q3} le premier quartile, la médiane et le troisième quartile, nous définissons les deux écarts suivants:

$$d_{Q1} = \text{Max} (E_M - E_{Q1}, |AE_M|),$$

$$d_{Q3} = \text{Max} (E_{Q3} - E_M, |AE_M|).$$

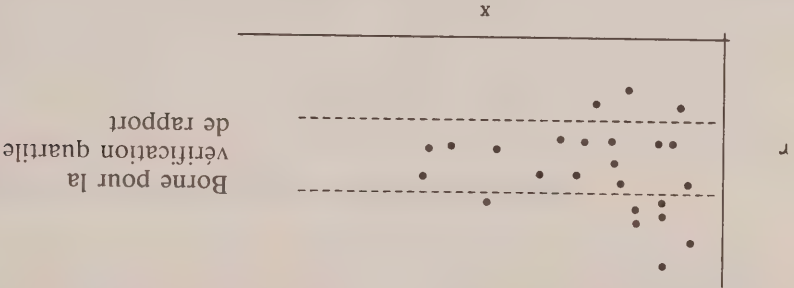
Toutes les unités dont l'effet E_i correspondant se situe à l'extérieur de l'intervalle $(E_M - Cd_{Q1}, E_M + Cd_{Q3})$ sont définies des observation aberrantes. Le terme AE_M sert à contourner la difficulté qui surgit lorsque $E_M - E_{Q1}$ ou $E_{Q3} - E_M$ sont très faibles. En d'autres mots, le cas où les effets E_i sont concentrés autour d'une seule valeur avec un ou deux faibles écarts peut produire de fausses observation aberrantes. Le paramètre C détermine la grandeur des intervalles d'acceptation. Le paramètre U détermine la forme de la courbe qui définit les bornes inférieure et supérieure. En augmentant la valeur de U , nous accordons par le fait même plus d'importance aux fluctuations des valeurs les plus fortes, à notre avis, le paramètre A devrait prendre la valeur 0.05 puisque celle-ci s'est avérée efficace dans la pratique.

2.1.3 Traitement des observations aberrantes

Une fois que l'on croit avoir décelé des observations aberrantes chez certaines unités, on les identifie comme telles et on signale leur existence aux enquêteurs. Ceux-ci doivent alors décider de la manière dont ces observations seront traitées, car leur existence peut être attribuable à plusieurs facteurs, par exemple une erreur de mesure, une mauvaise interprétation du questionnaire par l'unité répondante ou la variabilité inhérente de la population visée par l'enquête. S'il s'agit d'une erreur de mesure découlant d'une mauvaise transcription des données ou de réponses inexactes, un simple suivi permettra de corriger la majorité des erreurs. Lorsque des unités présentent une variabilité inhérente à cause d'une croissance rapide, les données fournies sont exactes mais elles se détachent trop des autres données contenues dans les tableaux récapitulatifs de l'enquête. Dans de telles circonstances, il faut appliquer des méthodes qui visent à minimiser l'effet des observations aberrantes. Parmi ces méthodes, notons celle proposée par Hidiroglou et Srinath (1981), qui permet de réduire le poids d'échantillonnage, et celle proposée par Ernst (1980), qui permet de modifier les valeurs proprement dites. Lorsque des unités fournissent des données non représentatives qui ne peuvent être vérifiées, on doit recourir à l'imputation. Les diverses solutions adoptées doivent être signalées au même titre que les observations aberrantes.

La procédure décrite ci-dessus est appliquée au dernier niveau de stratification de l'échantillon, c'est-à-dire au niveau des cellules. Si une cellule ne contient pas assez d'unités, elle doit être combinée à d'autres cellules qui présentent les mêmes caractéristiques.

L'utilisation de quartiles et d'interquartiles pour le calcul des bornes inférieure et supérieure, au lieu de la moyenne et de l'erreur type, a contribué à améliorer cette méthode. Il s'agit en l'occurrence de définir la borne inférieure $r_M - k D_{r_{Q1}}$ et la borne supérieure $r_M + k D_{r_{Q3}}$ où r_M est la médiane des rapports; $D_{r_{Q1}}$ est la distance entre le premier quartile et la médiane $D_{r_{Q3}}$ est la distance entre le troisième quartile et la médiane. Comme les queues de la distribution n'influent aucunement sur les quartiles, l'effet de masque se trouve largement atténué. Cette méthode présente toutefois deux inconvénients. Premièrement, il se peut, dans des circonstances très particulières, que la méthode ne permette pas de déceler des observations aberrantes à l'extrémité gauche de la distribution. Deuxièmement, cette méthode ne tient pas compte du fait que, dans la plupart des enquêtes-entreprises périodiques, la variabilité des rapports pour les petites entreprises est plus grande que la variabilité des rapports pour les grandes entreprises (Sugavanam, 1983). Ce fait est illustré dans le graphique ci-dessous.



Cet inconvénient se traduit par une suppréssion de petites unités et une sous-représentation de grandes unités dans la catégorie des unités à observations aberrantes. C'est ce qu'on appelle l'effet de masque par la taille.

2.1.2 Méthode proposée

Pour deux périodes t et $t + 1$, la tendance globale pour la paire d'éléments définie par

$$(x_i(t), x_i(t + 1)), i = 1, \dots, n$$

est

$$R = \sum_{i=1}^n x_i(t + 1) / \sum_{i=1}^n x_i(t).$$

Or, la fonction R peut s'exprimer comme suit:

$$R = \sum_{i=1}^n I_i r_i$$

où

$$I_i = x_i(t) / \sum_{i=1}^n x_i(t)$$

et

$$r_i = x_i(t + 1) / x_i(t).$$

I_i est une mesure de l'importance relative de la i ème unité parmi les n unités à la période t . Les tendances r_i doivent être transformées de façon qu'on puisse déceler les observations aberrantes aux deux extrémités de la distribution. Cette transformation est définie comme suit:

$$s_i = \begin{cases} r_i / r_M - 1, & \text{si } r_i \geq r_M \\ 1 - r_M / r_i, & \text{si } 0 < r_i < r_M \end{cases}$$

où r_M est la médiane des rapports.

périodes. D'autres auteurs ont proposé des définitions des observations aberrantes; ces définitions sont quelque peu vagues mais pratiques, comme en font foi les énoncés suivants.

Pour GRUBBS (1969), "une observation aberrante est une valeur qui semble s'écarter sensiblement des autres valeurs contenues dans l'échantillon" (Traduction)

Pour GUMBEL (1960), "les observations aberrantes sont des valeurs qui semblent trop élevées ou trop faibles par rapport au reste des observations." (Traduction)

Pour leur part, KENDALL et BUCKLAND (1957, p. 209), écrivent: "Dans un échantillon

de n observations, il se peut qu'un nombre limité de ces observations s'écarterent des autres à un point tel qu'il faille se demander si elles n'appartiennent pas à une autre population ou si la méthode d'échantillonnage n'est pas erronée. Ce genre de valeurs sont appelées des observations aberrantes. Il existe des tests qui permettent de déterminer si ce genre d'observations peuvent être assimilées au reste de l'échantillon." (Traduction)

2.1.1 Analyse de quelques méthodes actuellement utilisées

Dixon (1953), Grubbs (1969), Tietgen et Moore (1972) et Prescott (1978), pour ne nommer que ceux-là, ont proposé des méthodes permettant de repérer des valeurs aberrantes. La plupart de ces méthodes reposent sur les tests d'hypothèses. Dans les cas les plus simples, l'hypothèse nulle est que l'échantillon vient d'une distribution normale dont la moyenne et la variance sont inconnues, tandis que l'hypothèse alternative stipule qu'au moins une des observations provient d'une distribution différente. On peut déterminer les valeurs en pourcentage d'une fonction des observations selon l'hypothèse nulle et les comparer à des valeurs de cette même fonction calculées dans des applications particulières. Il est toutefois difficile d'appliquer ces méthodes à des données périodiques provenant de grandes enquêtes pour les raisons suivantes. Premièrement, l'hypothèse de la distribution normale des tendances d'une période à une autre peut ne pas être valide. Deuxièmement, ces méthodes classiques requièrent l'utilisation de tables permettant de déterminer les valeurs limites d'acceptation qui définissent les intervalles de rejet. Ces tables ne peuvent contenir toutes les valeurs nécessaires lorsqu'il faut tester plus de 100 observations. Troisièmement, pour appliquer ces méthodes, il faut prévoir le nombre d'observations aberrantes qui seront relevées. Quatrièmement, les étapes permettant d'isoler une à une les observations aberrantes exigent une bonne connaissance des données et des tests servant à déceler ces observations. La méthode que nous proposons dans la section 2.1.2 ne comporte aucun de ces inconvénients. Elle peut être appliquée facilement par ordinateur, elle n'exige pas l'hypothèse de la distribution normale et ne requiert pas de tables. Dans le cas qui nous intéresse, et étant donné les éléments des vecteurs $\hat{x}_i(t)$ et $\hat{y}_i(t+1)$, nous désignons $x_i(t)$ et $x_i(t+1)$ les réponses fournies par une unité donnée pour deux périodes consécutives, où $i = 1, \dots, n$. Nous définissons r_i le rapport entre les données de la période courante et celles de la période précédente. Nous appliquons alors la méthode connue sous le nom de méthode de contrôle par intervalle, qui consiste simplement à définir selon l'expérience des bornes inférieure et supérieure fixes qui serviront de point de comparaison. Les rapports qui tombent à l'extérieur de l'intervalle délimité par ces bornes sont considérés comme des valeurs aberrantes. L'une des principales faiblesses de cette méthode est de présenter une définition trop subjective de la valeur aberrante, qui fait abstraction de la distribution des rapports.

Il existe une méthode qui permet d'utiliser la distribution des rapports: la méthode de contrôle de l'inégalité de Chebychev. Cette méthode consiste à définir la borne inférieure $\bar{r} - k\sigma_r$ et la borne supérieure comme étant $\bar{r} + k\sigma_r$ où $\bar{r} = \sum_{i=1}^n r_i/n$ et $\sigma_r^2 = \sum_{i=1}^n (r_i - \bar{r})^2/(n-1)$. Cette méthode présente deux grandes lacunes. Premièrement, comme le choix de k est subjectif, la méthode peut être inefficace. Cet argument a été démontré par Wilkison (1982). Deuxièmement, les observations aberrantes de grande taille peuvent dissimuler des observations aberrantes de moindre importance. C'est ce qu'on appelle l'effet de masque.

2. CONTRÔLE DE DONNÉES PÉRIODIQUES

2.0 Contrôle de la cohérence

Pour une unité donnée i et une période t , définissons $\tilde{x}_i(t)$ comme le vecteur des données qui doivent être recueillies. Ce vecteur $\tilde{x}_i(t)$ peut être décomposé en une série de vecteurs élémentaires qui doivent tour à tour faire l'objet d'un contrôle et d'une imputation.

Ainsi, $\tilde{x}_i(t) = (\tilde{x}_{i(1)}(t), \dots, \tilde{x}_{i(p)}(t))$

où $\tilde{x}_{i(d)}'(t) = (x_{i(d)}^{I(d)}(t), \dots, x_{i(d)}^{K(d)}(t))$

pour $i=1, \dots, n; d=1, \dots, P; t=1, \dots, T$.

Pour chaque vecteur élémentaire $\tilde{x}_{i(d)}'(t)$, le contrôle de la cohérence peut être représenté comme suit:

$$\tilde{A}^{(d)}(\tilde{x}_{i(d)}'(t))' \leq (\tilde{c}^{(d)})'$$

où $\tilde{A}^{(d)}$ est une matrice ℓ_p par k_p qui représente les règles auxquelles doivent satisfaire les éléments du vecteur élémentaire $\tilde{x}_{i(d)}'(t)$, et $\tilde{c}^{(d)}$ est un vecteur 1 par ℓ_p qui représente les contraintes. Cette formule permet de définir des contrôles de cohérence aussi bien pour des variables qualitatives que pour des variables quantitatives. Dans le premier cas, les contrôles de cohérence des données peuvent servir à vérifier si les variables correspondent à des valeurs bien définies. Dans le deuxième cas, ils peuvent servir à vérifier si les valeurs de certaines variables ne sont pas supérieures (ou inférieures) à d'autres ou si une combinaison linéaire est supérieure, égale ou inférieure à une variable donnée.

2.1 Contrôle statistique

Comme les données sont recueillies périodiquement, il s'agit d'isoler les observations aberrantes contenues dans la série chronologique. Pour les besoins de notre étude, nous définissons une observation aberrante i comme étant une observation dont la tendance d'une période à une autre, pour des variables données du vecteur élémentaire $\tilde{x}_i(t)$, diffère sensiblement de la tendance globale correspondante des autres observations se rattachant aux mêmes sous-ensembles de la population. On peut également appliquer les contrôles statistiques à l'intérieur d'une même période en comparant les rapports de deux variables interdépendantes pour un sous-ensemble donné de la population. Dans le présent article, nous nous contenterons d'analyser l'application du contrôle statistique du point de vue de la tendance entre deux

Contrôle statistique et imputation dans les enquêtes-entreprises périodiques

M.A. HIDIROGLOU et J.-M. BERTHELOT¹

RÉSUMÉ

Dans toutes les enquêtes-entreprises mensuelles, trimestrielles ou annuelles, il est nécessaire de vérifier les données fournies par les unités répondantes et d'en imputer pour les unités non répondantes. Le présent document porte sur les diverses méthodes de contrôle et d'imputation des données. Le contrôle des données comprend le contrôle statistique et le contrôle de la cohérence. L'imputation est effectuée aussi bien pour les cas de non-réponse totale que pour ceux de non-réponse partielle.

MOTS CLÉS: Enquête périodique; contrôle statistique; non-réponse totale ou partielle; imputation.

1. INTRODUCTION

Les grands organismes comme Statistique Canada recueillent régulièrement des données au moyen d'enquêtes par sondage. Si ces données sont recueillies périodiquement auprès de la même unité d'échantillonnage, plusieurs possibilités peuvent être envisagées relativement à la cohérence (qualité) des données sur une période déterminée. En effet, l'unité d'échantillonnage peut fournir fidèlement les données sans qu'il y ait d'écart appréciable d'une période à l'autre. Les données peuvent aussi être déclarées fidèlement mais comporter des variations suspectes d'une période à l'autre. Ou encore, l'unité d'échantillonnage peut ne pas fournir tous les éléments d'information qui lui sont demandés : c'est ce qu'on appelle la non-réponse partielle. Enfin, l'unité d'échantillonnage peut répondre sporadiquement à l'enquête, négligeant par le fait même de fournir des données pour certaines périodes : c'est ce qu'on appelle la non-réponse totale. Ces cas peuvent se produire simultanément au cours d'une enquête périodique qui vise à recueillir des données auprès d'un grand nombre d'unités d'échantillonnage. Dans cet article, nous examinons le contrôle et l'imputation des données des unités d'échantillonnage qui sont contactées périodiquement par un organisme d'enquête. Les méthodes qui y sont analysées s'appliquent d'une façon générale à des données de nature diverse comprenant aussi bien des variables quantitatives que des variables qualitatives. Le contrôle englobe le contrôle de la cohérence des données et le contrôle statistique.

Pour ce qui a trait aux données quantitatives, le contrôle de la cohérence vise à assurer que les combinaisons linéaires des variables dans une période déterminée satisfont à des conditions précises. Dans le cas des données qualitatives, les contrôles de cohérence visent à vérifier si les variables correspondent à des valeurs bien définies.

Les contrôles statistiques servent à isoler les unités d'échantillonnage qui pourraient fournir des données quantitatives incohérentes d'une période à une autre ou à l'intérieur d'une même période. Les unités pour lesquelles il existera des valeurs anormalement élevées ou anormalement faibles seront désignées comme observations aberrantes. Il est très important de pouvoir repérer ce genre d'unités dans une enquête permanente et ce, pour deux raisons. Premièrement, elles ont une incidence sur les paramètres statistiques de l'ensemble de données, comme les totaux. Hidiroglou et Srinath (1981) ont analysé cet aspect de la question. Deuxièmement, comme l'imputation de données quantitatives pour les unités non répondantes

¹ M.A. Hidiroglou et J.-M. Berthelot, Division des méthodes d'enquêtes-entreprises, 11^e étage, Immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario), KIA 0T6.

$$\hat{\gamma} = (G'QG + \lambda I - G' \text{vec}(E)).$$

La valeur numérique de l'estimation de \mathcal{L} peut être déterminée à l'aide de l'équation de vraisemblance $G'H = 0$, où H est défini en (3). Au point de vue numérique, nous pouvons utiliser l'algorithme de Newton-Raphson défini à la section 3 en précisant toutefois que l'estimateur de $\hat{\gamma}$ à chaque itération est défini par

BIBLIOGRAPHIE

- CARTER, E.M. (1986). The analysis of a generalized multivariate linear model. Rapport technique, University of Guelph.
- DAVID, M., LITTLE, R.J., SAMUEL, M.E., et TRIEST, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.
- DEMPSTER, A.P., LAIRD, N.M., et RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 39, 1-38.
- DRAPER et SMITH (1981). *Applied Regression Analysis*. New York: Wiley.
- GREENLEES, W.S., REECE, J.S., et ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- HARTLEY, H.O., et HOCKING, R.R. (1971). The analysis of incomplete data (with discussion). *Biometrics*, 27, 783-823.
- HECKMAN, J.D. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimation for such models. *Annals of Economic and Social Measurements*, 5, 475-492.
- LITTLE, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- RUBIN, D.B., et SZATROWSKI, T.H. (1982). Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm. *Biometrika*, 69, 657-660.
- SRIVASTAVA, M.S. (1985). Multivariate data with missing observations. *Communications in Statistics Theory and Methods*, 14, 775-792.
- SRIVASTAVA, M.S., et WORSLEY, K.J. (1986). Likelihood ratio tests for a change in the multivariate mean. *Journal of the American Statistical Association*, 81, 199-204.
- WEI, L.J., et LACHIN, J.M. (1984). Two sample asymptotically distribution free tests for incomplete multivariate observations. *Journal of the American Statistical Association*, 79, 653-661.
- WOOLSON, R.F., et LEEPER, J.D. (1980). Growth curve analysis of complete and longitudinal data. *Journal of the American Statistical Association*, 9, 1491-1513.
- WOOLSON, R.F., LEEPER, J.D., et CLARKE, W.R. (1978). Analysis of incomplete data from longitudinal and mixed longitudinal studies. *Journal of the American Statistical Society, Ser. A*, 141, 242-252.

pour les N observations; par conséquent, les équations (2) et (3) permettent de déterminer l'estimateur de Σ ($\hat{\Sigma}$) pour le modèle à une population. L'équation (14) permet de déterminer la fonction du rapport de vraisemblance.

La méthode ci-dessus peut être modifiée. On pourrait, par exemple, imaginer les vecteurs \bar{y}_j ($j = 1, \dots, N$) comme des vecteurs de moyennes d'échantillon pour N périodes d'échantillonnage. Il est possible de déterminer plusieurs points d'inflexion en réappliquant la méthode à chaque groupe de données. Si nous avons 50 observations et le point d'inflexion se trouve à l'observation n° 20, nous appliquons successivement la méthode au groupe des observations 1 à 20 et au groupe des observations 21 à 50.

6. MATRICES DE VARIANCES-COVARIANCES STRUCTURÉES

Dans les études longitudinales, il se peut que les vecteurs d'erreurs ne soient pas arbitraires mais correspondent à un modèle de série chronologique. Si nous pouvons supposer l'existence d'un tel modèle, le nombre de paramètres à estimer s'en trouve réduit. Une série chronologique stationnaire supposerait que nous pouvons exprimer la matrice des covariances Σ comme suit,

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \dots & \rho_{p-1} \\ \rho_1 & 1 & \dots & \rho_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \dots & 1 \end{bmatrix} \quad (15)$$

Il est possible d'obtenir d'autres modèles. Nous pourrions structurer les coefficients de corrélation ρ_j . Par exemple, nous pourrions poser ρ_j égal à $\rho^{|j|}$. La méthode de Newton-Raphson pourrait servir à résoudre les équations de vraisemblance. Carter (1986) s'est penché sur le cas où la matrice de variances-covariances peut s'exprimer sous la forme $\text{vec}(\Sigma) = G\bar{\gamma}$ pour une matrice quelconque G . En posant $\gamma_i = \sigma^2 \rho_i$ pour $i = 1, \dots, p - 1$ et $\gamma_p = \sigma^2$, nous pouvons exprimer la matrice de variances-covariances pour la série chronologique stationnaire sous cette forme purement linéaire. Si, par exemple, $p = 3$, nous avons

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{21} \\ \sigma_{22} \\ \sigma_{23} \\ \sigma_{31} \\ \sigma_{32} \\ \sigma_{33} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}$$

observées en se servant de la méthode décrite à la section 2. Une fois cette opération terminée, on peut appliquer la méthode de la section 3.1 pour estimer les paramètres de régression et imputer les variables de réponses manquantes.

4.3 Test du rapport de vraisemblance.

Le test du rapport de vraisemblance permet de déterminer si les variables du modèle sont significatives. Pour tester l'hypothèse

$$H: \beta = \beta_1 F \text{ vs } A: \beta \neq \beta_1 F,$$

pour une matrice F de plein rang et de dimension $m \times q$ nous obtenons des estimations de Σ suivant l'hypothèse nulle (Σ) et l'hypothèse alternative (Σ). L'hypothèse nulle est rejetée au niveau de signification α si

$$\begin{aligned} & -2 \ln \lambda > \chi^2_{(q-m)p; \alpha}, \\ \text{ou} \quad & \lambda = \prod_{k=1}^K |B_k \tilde{\Sigma} B_k'|^{n_k/2} / |B_k \Sigma B_k'|^{n_k/2}. \end{aligned} \tag{14}$$

5. ESTIMATION D'UN POINT D'INFLEXION

Considérons une série d'observations y_j , $j = 1, \dots, N$, dont l'espérance mathématique est égale à $E(y_j) = \mu_j$. Srivastava et Worsley (1986) ont défini une méthode permettant d'estimer le point d'inflexion des vecteurs de moyennes μ_j . On suppose tout d'abord que l'inflexion se situe à un point quelconque r . Ensuite, on teste l'hypothèse suivante,

$$H: \bar{\mu}_1 = \dots = \bar{\mu}_N$$

$$\text{vs} \quad A: \bar{\mu}_1 = \dots = \bar{\mu}_r \neq \bar{\mu}_{r+1} = \dots = \bar{\mu}_N.$$

On calcule par la suite la valeur λ_r , de la fonction du rapport de vraisemblance pour $r = 1, \dots, N - 1$. Le point d'inflexion estimé correspond à la valeur de r pour laquelle λ_r est maximale.

L'existence de données incomplètes ne pose pas de problème dans l'estimation du point d'inflexion. Le modèle linéaire est défini comme dans le cas des données complètes et, par conséquent, les observations sont réparties en K sous-ensembles. Supposons que la portion observée de y_j est désignée \bar{z}_{ki} . Alors, selon l'hypothèse alternative pour une valeur donnée de r , Σ , l'estimateur de Σ , est obtenu de l'équation (3) pour le modèle de régression défini dans les équations (9) à (12), où la matrice des paramètres β est définie par

$$\beta = (\bar{\mu}_1, \bar{\mu}_2)$$

et la matrice de base pour le k -ième sous-ensemble est définie par

$$A_k = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \end{bmatrix}.$$

Si l'observation \bar{z}_{ki} correspond au vecteur y_j et $j \leq r$, on trouve un 1 à l'intersection de la première ligne et de la i -ième colonne de A_k et, autrement, on trouve un zéro. Selon l'hypothèse nulle, on considère que le vecteur de moyennes de la population est le même

L'opération est ensuite répétée. Il est avantageux de recourir à l'algorithme EM lorsqu'il existe des solutions complètes simples pour les équations de vraisemblance en situation de données complètes. S'il faut recourir à la méthode de Newton-Raphson pour résoudre les équations de vraisemblance pour données complètes, l'algorithme EM s'avère peu utile.

4. MODÈLE DE RÉGRESSION

4.1 Variables de réponses incomplètes.

Le modèle étudié à la section 2 peut également faire l'objet d'une analyse de régression. Nous répartissons de nouveau les données en K sous-ensembles. Puis nous définissons le modèle de régression suivant:

$$Z_k = B_k' \beta A_k + \epsilon_k, \text{ pour } k = 1, \dots, K,$$

où Z_k est la matrice $p_k \times n_k$ des valeurs observées, β est une matrice $p \times q$ de paramètres inconnus, B_k est la matrice définie à la section 2, A_k est la matrice de base de la matrice Z_k et les colonnes de valeurs ϵ_k sont distribuées indépendamment avec une moyenne 0 et une matrice de variances-covariances $B_k' \Sigma B_k$. Pour une matrice Σ donnée, l'estimateur de β par les moindres carrés peut être formulé de la façon suivante (Carter 1986):

$$\text{vec}(\hat{\beta}) = P^{-1} \text{vec}(E), \tag{9}$$

où

$$P = \sum_{k=1}^K n_k B_k' (B_k' \Sigma B_k)^{-1} B_k \otimes A_k A_k', \tag{10}$$

$$E = \sum_{k=1}^K B_k' (B_k' \Sigma B_k)^{-1} Z_k A_k'. \tag{11}$$

L'estimateur du maximum de vraisemblance de Σ est défini comme en (3), sauf que dans ce cas-ci,

$$V_k = [Z_k - B_k \beta A_k][Z_k - B_k \beta A_k]'. \tag{12}$$

La distribution asymptotique de $\hat{\beta}$ peut s'exprimer sous la forme

$$\text{vec}(\hat{\beta}) \sim N_{pq}(\text{vec}(\beta), P^{-1}). \tag{13}$$

Cette distribution de même que la fonction du rapport de vraisemblance définie par Srivastava (1985) permettent de faire de l'inférence statistique sur les paramètres de régression.

4.2 Variables explicatives incomplètes

Dans la section 3.1, nous avons supposé que les matrices de base étaient entièrement connues. Il peut arriver que les variables explicatives soient elles aussi incomplètes. Si ces variables sont aléatoires, on peut tout d'abord imputer les valeurs manquantes à partir des données

où $A \otimes B$ désigne le produit de Kronecker de deux matrices A et B défini par $A \otimes B = (a_{ij}B)$,

$$D_k = B_k (B_k \Sigma_0 B_k')^{-1} B_k,$$

et

$$F_k = B_k (B_k \Sigma_0 B_k')^{-1} V_k (B_k \Sigma_0 B_k')^{-1} B_k.$$

Pour toute matrice $A = (\bar{a}_1, \dots, \bar{a}_q)'$, posons $\text{vec}(A) = (\bar{a}_1', \dots, \bar{a}_q')$. L'équation (3) peut donc être réexprimée approximativement comme suit,

$$E = \sum_k^K (D_k - F_k). \\ \bar{Q} \text{ vec}(\Lambda) = \text{vec}(E),$$

où

Pour faire en sorte que \bar{Q} soit non-singulière, nous exprimerons la solution de $\text{vec}(\Lambda)$ comme suit,

$$(8) \qquad \text{vec}(\Lambda) = (\bar{Q} + \lambda I)^{-1} \text{vec}(E),$$

où λ peut varier selon l'algorithme mais a une valeur initiale très faible. Pour une valeur donnée de Σ , on détermine $\bar{\mu}$ à l'aide de l'équation (2), puis on calcule une valeur de Λ au moyen de l'équation (8) afin d'obtenir une estimation révisée de Σ . On répète ensuite l'opération jusqu'à ce qu'on obtienne le degré de convergence voulu.

La méthode décrite ci-dessus peut être appliquée aux matrices de covariances structurées, plus complexes; il faut toutefois pour cela inverser $\bar{Q} + \lambda I$. Lorsque l'il y a un grand nombre de variables, cette matrice prend des dimensions très appréciables. Il est alors préférable de résoudre l'équation (3) au moyen de l'algorithme EM. Comme il s'agit encore d'une méthode itérative, les calculs doivent être faits à l'aide des estimations révisées de $\bar{\mu}$ et de Σ à chaque itération. Pour une forme particulière de Σ , par exemple Σ_0 , définissons le vecteur complet estimé $\bar{y}_{kj} = B_k' \bar{z}_{kj} + C_k' \bar{\mu}_{kj}$, où la valeur manquante estimée $\bar{\mu}_{kj}$ est définie par (6). Alors,

Définissons la matrice V comme suit,

$$V = \sum_K^{n_K} \sum_{j=1}^{k=1} (\bar{y}_{kj} - \bar{\mu})(\bar{y}_{kj} - \bar{\mu})'.$$

L'estimation révisée de Σ est alors définie par l'équation suivante

$$\bar{\Sigma} = (1/N) [V + \sum_K^{n_K} C_k' H C_k].$$

où H_k est la variance conditionnelle des données incomplètes étant donné les données observées pour la k -ième sous-ensemble et est définie comme

$$H_k = C_k \Sigma C_k' - (C_k \Sigma B_k') (B_k \Sigma B_k')^{-1} (B_k \Sigma C_k').$$

2.5 Imputation

L'imputation des données manquantes peut se faire à l'aide de la distribution conditionnelle des données non observées, en fonction des renseignements observés. Ainsi, définissons les matrices C_k , $k = 1, \dots, K$, comme les compléments de B_k . En d'autres termes, pour une matrice B_k de dimension $p_k \times p$ composée de (1) à l'intersection de la s -ième ligne et de la i_s -ième colonne $s = 1, \dots, p_k$ et de (0) aux autres intersections, on définit la matrice C_k comme une matrice $(p - p_k) \times p$ composée des nombres un et zéro, les premiers se trouvant à l'intersection de la i_t -ième ligne et de la i_t -ième colonne et les seconds occupant le reste de la matrice, étant donné $i_t \neq i_s$ pour toutes les valeurs de $t = 1, \dots, (p - p_k)$ et toutes les valeurs de $s = 1, \dots, p_k$. Si le vecteur de réponses \bar{y}_{kj} correspond à la j -ième observation du sous-ensemble k , le vecteur de réponses observé est $\bar{z}_{kj} = B_k \bar{y}_{kj}$ et le vecteur non observé est $\bar{u}_{kj} = C_k \bar{y}_{kj}$. La valeur estimée du vecteur manquant est définie par l'équation suivante

$$\bar{u}_{kj} = (C_k \bar{u} + [C_k \hat{B}_k] [B_k \hat{B}_k]^{-1} (\bar{z}_{kj} - B_k \bar{u})) \quad (6)$$

Il convient de souligner que l'estimation ci-dessus ne renferme pas d'erreur aléatoire. Si l'on devait utiliser ultérieurement ces valeurs imputées ainsi que les données observées qui les accompagnent comme s'il s'agissait d'une série de données complètes, la matrice de variances-covariances résiduelles estimée serait alors trop petite. Il serait possible de surmonter cette difficulté en associant une valeur résiduelle convenable $\bar{\epsilon}$ à l'estimation \bar{u}_{kj} . Si le premier sous-ensemble de données complètes est suffisamment grand, on peut prélever aléatoirement les vecteurs des valeurs résiduelles se rattachant aux observations manquantes du sous-ensemble k parmi la série de valeurs

$$(C_k \bar{y}_{1i} - C_k \bar{u}) - [C_k \hat{B}_k] [B_k \hat{B}_k]^{-1} (B_k \bar{y}_{1i} - B_k \bar{u}) \quad \text{pour } i = 1, \dots, n_1. \quad (7)$$

Exemple 1 (suite):

Le Tableau 1 contient les séries de données complètes, y compris les valeurs imputées fondées sur les expressions (6) et (7), pour les sous-ensembles 2 à 8, les chiffres entre parenthèses étant les valeurs imputées.

3. MÉTHODES DE CALCUL

Les équations (2) et (3) peuvent être résolues par itération. Carter (1986) propose à cet égard un algorithme qui est une combinaison de la méthode de Newton-Raphson et de la méthode de la plus forte pente ascendante pour un cas général qui comprend uniquement des moyennes et des covariances à relation linéaire. Cet algorithme peut être décrit comme suit. Supposons que nous choisissons initialement une matrice Σ , soit Σ_0 , de telle sorte que

$$\Sigma = \Sigma_0 + \Lambda$$

constitue une solution. Nous faisons alors la substitution nécessaire dans l'équation (3) et nous développons ensuite celle-ci en une série comportant uniquement les termes linéaires de Λ . Nous obtenons alors la solution approximative suivante pour Λ . Définissons

$$\bar{Q} = \sum_{k=1}^K (D_k \otimes D_k - D_k \otimes F_k - F_k \otimes D_k),$$

Toutefois, comme la matrice de variances-covariances doit être définie positive, toute expression qui est minimisée doit avoir une solution définie positive. Si un des groupes renferme des données complètes, l'expression (4) aura une solution indéfinie pour toute matrice singulière Σ ; en conséquence, cette expression aura un minimum dans l'espace des matrices définies positives. Le même raisonnement s'applique aux estimateurs du maximum de vraisemblance.

2.3 Distribution asymptotique de $\bar{\mu}$.

L'équation (2) nous permet de déduire que l'estimateur $\hat{\mu}$ admet asymptotiquement une loi normale de moyenne $\bar{\mu}$ et de matrice de variances-covariances

$$P = I \sum_{k=1}^K n_k B_k (B_k \Sigma B_k)^{-1} B_k \tag{5}$$

laquelle peut être estimée au moyen de P en substituant $\hat{\Sigma}$ à Σ . La théorie asymptotique permet de tester certaines hypothèses et de déterminer des intervalles de confiance pour $\bar{\mu}$ ou pour des combinaisons linéaires de $\bar{\mu}$. Par ailleurs, les tests du rapport de vraisemblance définis par Srivastava (1985) peuvent servir à tester l'hypothèse $H: \bar{\mu} = \bar{0}$ par rapport à l'hypothèse alternative $A: \bar{\mu} \neq \bar{0}$. Suivant le test du rapport de vraisemblance, l'hypothèse nulle H est rejetée si

$$\lambda = \prod [|B_k \hat{\Sigma} B_k| / |B_k \hat{\Sigma} B_k|]^{n_k/2} > \chi^2_{2, \alpha}$$

où $\hat{\Sigma}$ est l'EMV de Σ suivant H et $\chi^2_{2, \alpha}$ est la limite supérieure (pour un $100\alpha\%$ donné) d'une distribution de chi-carré avec p degrés de liberté.

2.4 Estimateurs du maximum de vraisemblance pour l'exemple 1

Les estimations les plus vraisemblables pour l'exemple 1 sont les suivantes:

$$\bar{\mu} = \begin{pmatrix} 226.82 \\ 249.78 \\ 252.02 \\ 255.15 \\ 255.22 \end{pmatrix} \text{ et } \hat{\Sigma} = \begin{pmatrix} 1809 & 1220 & 1033 & 873 & 913 \\ 1220 & 1642 & 992 & 1017 & 1121 \\ 1033 & 992 & 1438 & 718 & 1189 \\ 873 & 1017 & 718 & 1233 & 915 \\ 913 & 1121 & 1189 & 915 & 2508 \end{pmatrix}.$$

La matrice de variances-covariances estimée se rattachant à l'estimation du vecteur de moyennes est

$$P^{-1} = \begin{pmatrix} 28.05 & 18.78 & 15.96 & 13.46 & 14.08 \\ 18.78 & 25.67 & 15.42 & 15.84 & 17.51 \\ 15.96 & 15.42 & 24.19 & 11.24 & 19.31 \\ 13.46 & 15.84 & 11.24 & 23.33 & 15.38 \\ 14.08 & 17.51 & 19.31 & 15.38 & 54.77 \end{pmatrix}.$$

À l'aide de la distribution asymptotique des estimateurs définie dans la section 2.3, il est possible de faire de l'inférence sur $\bar{\mu}$.

2.2 Estimation du vecteur des moyennes et de la matrice de variances-covariances de la population.

Pour chacun des K sous-ensembles, définissons la moyenne de l'échantillon

$$\bar{z}_k = (n_k)^{-1} \sum_{j=1}^{n_k} z_{kj}$$

Alors

$$E(\bar{z}_k) = B_{k\mu},$$

$$\text{cov}(\bar{z}_k) = n_k^{-1} (B_k \Sigma B_k'),$$

et les valeurs \bar{z}_k sont distribuées indépendamment pour $k = 1, \dots, K$. Par la théorie des moindres carrés, nous minimisons

$$\sum_{k=1}^K \text{tr } n_k (B_k \Sigma B_k')^{-1} [\bar{z}_k - B_{k\mu}]' [\bar{z}_k - B_{k\mu}].$$

Pour une valeur donnée de Σ , nous avons la solution

$$\bar{\mu} = \left[\sum_{k=1}^K n_k B_k' (B_k \Sigma B_k')^{-1} B_k \right]^{-1} \left[\sum_{k=1}^K n_k B_k' (B_k \Sigma B_k')^{-1} \bar{z}_k \right]. \tag{2}$$

Si l'on suppose une distribution normale, l'estimateur par les moindres carrés est aussi l'estimateur du maximum de vraisemblance. Little (1982) propose de résoudre ce problème par l'algorithme EM (espérance mathématique) et soutient qu'il n'est pas nécessaire de supposer une distribution normale. En d'autres termes, on peut définir les estimateurs de $\bar{\mu}$ et de Σ comme la solution des équations normales de vraisemblance même si la population à l'étude n'est pas distribuée suivant une loi normale. Ainsi, on ne peut plus considérer ces estimateurs comme des estimateurs du maximum de vraisemblance mais uniquement comme des estimateurs heuristiques qui ne sont convergents qu'à certaines conditions générales. Cependant, si l'on ne suppose pas de distribution normale, il n'y a aucune raison de maximiser les équations normales de vraisemblance pour obtenir des estimateurs. À la fin de cette section, nous définissons un estimateur heuristique pour Σ . L'estimateur du maximum de vraisemblance pour Σ , en supposant une distribution normale, est défini par Srivastava (1985) comme étant la solution de l'équation suivante:

$$H = \sum_{k=1}^K n_k B_k' (B_k \Sigma B_k')^{-1} B_k - \sum_{k=1}^K B_k' (B_k \Sigma B_k')^{-1} V_k (B_k \Sigma B_k')^{-1} B_k = 0, \tag{3}$$

où

$$V_k = (\bar{z}_{k1} - B_{k\mu}, \dots, \bar{z}_{kn_k} - B_{k\mu})' (\bar{z}_{k1} - B_{k\mu}, \dots, \bar{z}_{kn_k} - B_{k\mu})'.$$

On donne à la section 3 des méthodes de calcul permettant de résoudre les équations (2) et (3).

Note: Il est possible de déterminer des estimateurs heuristiques pour la matrice de covariances sans supposer une distribution normale. Nous pouvons, par exemple, définir Σ comme la valeur de Σ qui minimise

$$\sum_{k=1}^K n_k^{-1} \text{tr} [(B_k \Sigma B_k')^{-1} V_k - n_k I_k]^2. \tag{4}$$

Tableau I

Taux de cholestérol observés et valeurs imputées

Variable									
1	2	3	4	5					
224	273	242	274	(231)	Sous-ensemble 2: $n_2 = 7$				
231	252	267	299	(233)					
268	296	314	330	(303)					
284	288	268	261	(300)					
217	231	276	257	(238)					
209	200	269	233	(323)					
200	261	264	300	(279)					
193	189	(257)	232	211	Sous-ensemble 3: $n_3 = 1$				
201	219	220	(231)	(172)	Sous-ensemble 4: $n_4 = 12$				
202	186	253	(245)	(328)					
209	207	167	(208)	(194)					
212	253	225	(157)	(194)					
276	326	304	(300)	(376)					
163	179	199	(211)	(224)					
239	243	265	(238)	(246)					
204	203	198	(234)	(171)					
247	211	225	(224)	(215)					
195	250	272	(265)	(231)					
228	228	279	(276)	(259)					
290	264	260	(249)	(325)					
227	247	(215)	(267)	220	Sous-ensemble 5: $n_5 = 1$				
250	269	(327)	(250)	(295)	Sous-ensemble 6: $n_6 = 5$				
175	214	(250)	(210)	(210)					
260	268	(327)	(248)	(321)					
197	218	(235)	(251)	(258)					
248	262	(286)	(251)	(271)					
193	(209)	(219)	(230)	(255)	Sous-ensemble 7: $n_7 = 2$				
256	(277)	(294)	(260)	(281)					
(284)	327	(287)	(336)	(309)	Sous-ensemble 8: $n_8 = 1$				

Note: La taille globale de l'échantillon est $N = 65$.

l'échantillon est $N = n_1 + \dots + n_K$. Si le k -ième sous-ensemble renferme p_k caractères i_1, \dots, i_k , la matrice B_k sera alors une matrice $p_k \times p$ constituée des nombres un et zéro, les premiers se trouvant à l'intersection de la i_s -ième ligne et de la i_s -ième colonne pour $s = 1, \dots, p_k$ et les seconds occupant le reste de la matrice. Compte tenu de cette notation, nous pouvons exprimer les vecteurs de réponses observés comme suit:

D'où,

$$\bar{z}_{kj} = B_k \bar{y}_{kj}, j = 1, \dots, n_k, k = 1, \dots, K.$$

$$E(\bar{z}_{kj}) = B_k \bar{\mu},$$

et

$$\text{cov}(\bar{z}_{kj}) = B_k \Sigma B_k', j = 1, \dots, n_k \text{ et } k = 1, \dots, K.$$

Exemple 1: (données)

Wei et Lachin (1984) ont établi les taux de cholestérol chez un groupe de patients à différentes périodes; après un relevé initial, les taux ont été recalculés après 6 mois, 12 mois, 20 mois et 24 mois. Pour des raisons qui n'ont rien à voir avec la variable de réponse, certaines observations étaient incomplètes. Les données peuvent être réparties en $K = 8$ sous-ensembles. Pour le premier groupe de données complètes, le vecteur de moyennes de l'échantillon et la matrice de variances-covariances fondée sur 36 observations sont les suivantes:

$$\bar{z}_1 = \begin{pmatrix} 226.6 \\ 249.6 \\ 252.6 \\ 253.1 \\ 256.7 \end{pmatrix} \quad S_1 = \begin{pmatrix} 1964 & 1301 & 1151 & 960 & 1008 \\ 1301 & 1715 & 1109 & 1023 & 1199 \\ 1151 & 1109 & 1554 & 697 & 1266 \\ 960 & 1023 & 697 & 1148 & 667 \\ 1008 & 1199 & 1266 & 667 & 2546 \end{pmatrix}$$

Les données des autres sous-ensembles figurent dans le tableau 1, les chiffres entre parenthèses étant des valeurs imputées.

Les matrices qui définissent le modèle pour les valeurs observées sont:

$$B_1 = I_5 \quad B_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad B_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad B_4 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad B_5 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad B_6 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad B_7 = (1 \ 0 \ 0 \ 0 \ 0) \quad B_8 = (0 \ 1 \ 0 \ 0 \ 0).$$

Maintenant que nous avons défini le modèle, nous pouvons procéder à l'estimation des paramètres et à l'imputation des données manquantes.

Lorsque la non-réponse n'est pas aléatoire, il est aussi possible de tenir compte du caractère non aléatoire des données incomplètes en employant un nombre suffisant de variables explicatives dans le modèle de régression et en appliquant quelques-unes des techniques prévues dans la méthode "hot deck", comme l'ont fait David et coll. (1986) pour un modèle à une variable. Dans la section 2.5, par exemple, on définit une méthode d'imputation des valeurs manquantes.

Au cours de cet exercice, nous mettrons au point une méthode qui permettra de vérifier s'il y a eu une évolution dans la réponse. Les modèles utilisés pourront également être modifiés de manière à inclure des matrices de variances-covariances résiduelles, qui sont structurées par l'application d'une série chronologique aux variables de réponse. Dans cet article on suppose l'échantillonnage aléatoire simple, une distribution normale des données et qu'on peut considérer la non-réponse comme étant répartie aléatoirement. Si on ne suppose pas de distribution normale, les estimateurs ne sont plus les estimateurs du maximum de vraisemblance mais ils demeurent de bons estimateurs heuristiques.

Dans la section suivante, nous décrivons le modèle à un échantillon.

2. MODÈLE À UN ÉCHANTILLON

2.1 Définition

Nous examinons tout d'abord le modèle de données manquantes à deux variables; cette étude sert d'introduction à la définition du modèle général, qui vient plus loin. Soit $\bar{y} = (y_1, y_2)'$ un vecteur aléatoire à deux variables avec un vecteur de moyennes $\bar{\mu}$ et une matrice de variances-covariances Σ . Sans limiter la généralité de notre analyse, nous pouvons définir les données manquantes dans le modèle à deux variables de la façon suivante:

(1)
$$y_{11}, \dots, y_{1n_1}, y_{1,n_1+1}, \dots, y_{1,n_1+n_2} \text{ -----}$$
$$y_{21}, \dots, y_{2n_1}, \text{ ----- } y_{2,n_1+n_2+1}, \dots, y_{2,n_1+n_2+n_3}$$

Il y a donc n_1 couples d'observations sans données manquantes, n_2 observations relatives à y_1 pour lesquelles l'observation correspondante pour y_2 est manquante, et n_3 observations relatives à y_2 pour lesquelles l'observation correspondante pour y_1 est manquante. Ainsi, $N = n_1 + n_2 + n_3$ observations sont réparties en trois sous-ensembles. Si la série de données complètes était désignée y_1, \dots, y_N , les données observées pourraient être définies

$$\bar{z}_{1j} = B_1 \bar{y}_j = \bar{y}_j, \text{ pour } j = 1, \dots, n_1,$$
$$z_{2j} = B_2 \bar{y}_j = y_{1j}, \text{ pour } j = n_1 + 1, \dots, n_1 + n_2,$$
$$z_{3j} = B_3 \bar{y}_j = y_{2j}, \text{ pour } j = n_1 + n_2 + 1, \dots, n_1 + n_2 + n_3,$$

où $B_1 = I_2$, la matrice unité, $B_2 = (1, 0)$ et $B_3 = (0, 1)$.

En ce qui concerne le modèle général à plusieurs variables (à un échantillon), les données sont réparties en K sous-ensembles contenant respectivement n_1, \dots, n_K observations. Soulignons que le nombre maximum de groupes est $2^p - 1$. De plus, la taille globale de

Application de la méthode du maximum de vraisemblance au traitement de la non-réponse dans les enquêtes par sondage

M.S. SRIVASTAVA et E.M. CARTER¹

RÉSUMÉ

Des réponses incomplètes rendent difficile l'analyse de données d'enquête. En appliquant la méthode du maximum de vraisemblance, il est possible d'obtenir des estimateurs pour les paramètres à l'étude et d'effectuer certains tests statistiques. Dans le présent document, nous définissons les estimateurs du maximum de vraisemblance pour le cas où la non-réponse est considérée comme étant répartie aléatoirement. Nous examinons une méthode d'imputation des valeurs manquantes ainsi que le problème de l'estimation des points d'inflexion pour la moyenne. Nous tentons également d'étendre les résultats de notre analyse à des covariances structurées et au cas où la non-réponse n'est pas aléatoire.

MOTS CLÉS: Réponses incomplètes; non-réponse aléatoire; méthode du maximum de vraisemblance; imputation.

1. INTRODUCTION

Les exemples de non-réponse dans les enquêtes par sondage sont nombreux. Divers auteurs spécialisés ont tenté avec plus ou moins de succès de résoudre la question. L'efficacité d'une méthode particulière dépend de la complexité du problème. Par exemple, lorsque la non-réponse n'est pas aléatoire, le problème est loin d'être résolu. Les travaux de Heckman (1976) et de Greenlees et coll. (1982) notamment attribuent principalement cette complexité à une mauvaise définition des modèles. De même, la méthode "hot deck" a été vivement critiquée par divers auteurs. Cependant, lorsque la taille de l'échantillon est grande, cette méthode d'imputation est aussi efficace qu'une méthode de régression soigneusement définie; c'est effectivement le cas pour l'imputation du revenu dans la Current Population Survey (CPS). À ce sujet, voir David, Little, Samuël et Triest (1986).

La méthode de régression est fondée sur l'hypothèse selon laquelle la non-réponse est aléatoire et, contrairement à la méthode "hot deck", elle n'exige pas l'utilisation de données complètes provenant d'enquêtes antérieures, des données complètes étant de toute façon difficiles à obtenir. Il semble donc qu'une méthode de régression soigneusement élaborée peut être d'une grande utilité.

Dans notre exposé, nous examinons les cas où la non-réponse est aléatoire, ce qui se produit assez souvent. Dans un échantillonnage répété, par exemple, on débute avec un certain nombre de personnes auprès desquelles certains renseignements doivent être recueillis pendant une période déterminée. À la fin de cette période, une partie des répondants sont retirés de l'échantillon et remplacés par d'autres. On procède de cette façon jusqu'à ce que l'enquête soit terminée. Woolson, Leeper et Clarke (1978) et Wolson et Leeper (1980) examinent des cas semblables.

¹ M.S. Srivastava, Département de statistiques, University of Toronto, Toronto (Ontario), Canada, M5S 1A1, et E.M. Carter, Département de mathématiques, University of Guelph, Guelph (Ontario), Canada, N1G 2W1.

employée est très valable et qu'elle devrait être incorporée dans un nouveau système. Cependant, il se peut que le logiciel PSTAT ne constitue plus un logiciel approprié. Le système NEIS a été utilisé, dans un cadre d'exploitation, lors de l'Enquête de 1981 sur l'utilisation de l'énergie dans les fermes. La méthodologie a aussi été employée lors de l'élaboration du système de dépouillement du Recensement de l'agriculture de 1981.

Le système NEIS, comme le système CAN-EDIT, utilise la méthode de l'imputation par enregistrement donneur avec des variables d'appariement déterminées automatiquement au moyen de la procédure décrite dans la sous-section 3.2. Cependant, comme nous l'avons expliqué dans cette sous-section, quand les variables d'appariement sont déterminées de cette façon pour des données numériques, la procédure d'imputation ne produit pas toujours un enregistrement épuré. La stratégie adoptée pour réduire l'importance de ce problème consiste à choisir les enregistrements donneurs les plus rapprochés. Si l'enregistrement donneur le plus rapproché ne permet pas d'imputer des valeurs qui sont acceptées aux contrôles, c'est alors l'enregistrement donneur le plus rapproché suivant qui est considéré et ainsi de suite.

Le système NEIS ne permet pas à l'utilisateur de choisir une fonction de transformation ou une fonction de distance. Il utilise la transformée des valeurs de rang et la norme L_∞ pondérée pour le calcul des distances.

5. CONCLUSION

Les suggestions présentées ici offrent un choix considérable à tout utilisateur d'un système généralisé de contrôle et d'imputation. Comme nous l'avons dit, ces suggestions n'excluent pas la possibilité d'autres méthodes. Cependant, nous pensons qu'un système élaboré avec ces éléments répondrait aux besoins d'un grand nombre d'utilisateurs. L'expérience des auteurs permet d'affirmer que la puissance et l'utilité ultimes d'un tel système ne sont pas évidentes avant qu'on ne commence à l'utiliser. À mesure que les essais se déroulent, il devient apparent qu'un tel système comporte plus de possibilités et qu'il peut être augmenté plus qu'on ne le pensait d'abord.

BIBLIOGRAPHIE

- FELLEGI, I.P., et HOLT, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.
- KALTON, G., et KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 22-31.
- KALTON, G., et KISH, L. (1981). Two efficient random imputation procedures. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 146-151.
- LITTLE, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- MADOW, L.H., et MADOW, W.G. (1978). On link relative estimators. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 534-539.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SANDE, G. (1976). Searching for numerically matched records. Rapport technique, Division des méthodes d'enquêtes - entreprises, Statistique Canada.
- SANDE, G. (1979). *The Numerical Edit and Imputation Subsystem for PSTAT - A User's Guide*. Sous-division de la recherche et des systèmes généraux, Statistique Canada.

méthode est appelée lissage exponentiel. Il s'agit d'une technique de prévision utilisée couramment en économétrie. Un paramètre doit être précisé par l'utilisateur. Ce paramètre permet de modifier l'apport relatif des diverses valeurs des données. Algébriquement, l'estimateur est donné par la formule suivante:

$$y_t^u = \frac{1 - A^t}{1 - A^T} \sum_{r=0}^{T-1} A^r y_{t+(T-r-1)}$$

où $0 < A < 1$, est précisé à l'avance.

Plus A est près de zéro, plus le poids des données récentes est important. Si $t = 1$, cela revient à imputer la valeur déclarée lors de l'enquête précédente.

4. TRAVAUX ANTÉRIEURS EFFECTUÉS À STATISTIQUE CANADA

Statistique Canada a essayé, dans le passé, d'élaborer un système généralisé de contrôle et d'imputation. Deux de ces tentatives seront soulignées car elles forment la base du présent document. Ce sont le système CAN-EDIT et le système de contrôle et d'imputation des données numériques (NEIS).

4.1 CAN-EDIT

Le système CAN-EDIT lui-même n'est pas un système complètement général, mais la méthode qu'il utilise l'est. Le système est basé sur le document de Fellegi et Holt (1976) sur l'imputation pour les données catégorielles. Il a été élaboré pour dépouiller les données du recensement de la population et du logement effectués en 1976 et 1981 au Canada. La méthode utilisée dans le système CAN-EDIT est celle de l'imputation par enregistrement donneur. Les variables d'appariement ont été déterminées automatiquement au moyen des procédures décrites dans la sous-section 3.2. Le système CAN-EDIT emploie ce qui est appelé l'imputation primaire et l'imputation secondaire. Si l'imputation relative à un enregistrement receveur ne peut être effectuée par l'imputation primaire, cet enregistrement passe à l'imputation secondaire.

Dans l'imputation primaire, toutes les valeurs imputées sont tirées du même enregistrement donneur. Les variables d'appariement ont été déterminées d'après toutes les variables à imputer. Un enregistrement est rejeté lors de l'imputation primaire si aucun enregistrement donneur ne possède des valeurs identiques à celles des variables d'appariement. Dans l'imputation secondaire, chacune des variables à imputer est traitée indépendamment et séquentiellement. La procédure utilisée pour déterminer les variables d'appariement est la même. Cependant, comme on ne considère qu'une variable à la fois, le nombre de variables d'appariement sera en général moindre que lors de l'imputation primaire. (Il ne peut y en avoir plus, mais il pourrait y en avoir autant.) Cela signifie que la population d'enregistrements donneurs potentiels est plus grande. L'imputation secondaire présente quelques désavantages par rapport à l'imputation primaire. Tout d'abord, il est possible de choisir comme variable d'appariement, une variable qui doit être soumise à l'imputation. Il n'existe pas alors de valeur avec laquelle effectuer un appariement. Deuxièmement, cette méthode n'utilise pas les distributions conjointes des variables. Les valeurs imputées pour les deux variables peuvent être acceptées aux contrôles, chacune peut être une valeur réellement valide, mais il se peut qu'elles ne se trouvent que rarement combinées dans la population.

4.2 Système de contrôle et d'imputation des données numériques (NEIS)

Le système NEIS est le premier prototype d'un système généralisé de C&I pour les données numériques. Ce système a été écrit comme un ensemble de modules du progiciel statistique PSTAT. Aucun autre prototype n'a été élaboré par la suite. Ce système est décrit dans l'ouvrage de Gordon Sande (1979), qui l'a élaboré. Nous pensons que la méthodologie

Il est intéressant de faire ressortir les différences dans les estimateurs quant on fixe tous les éléments de classification sauf un. Par exemple, la différence entre les estimateurs 1 et 2 n'est due qu'à la différence dans le choix du groupe d'imputation, ce qui est aussi le cas pour les estimateurs 3 et 4 et pour les estimateurs 5 et 6. La différence entre les estimateurs 1 et 7 n'est due qu'au choix du modèle. Cela vaut aussi pour les estimateurs 3 et 5 et pour les estimateurs 4 et 6. Il faut aussi remarquer que les estimateurs 4 et 6 sont ceux utilisés lors de l'imputation par enregistrement donneur, sujet traité dans la sous-section 3.2.

3.4 Autres estimateurs d'imputation

Le choix des techniques d'imputation dépend des hypothèses faites par l'utilisateur à propos de l'ensemble des non-répondants. Quand on utilise l'imputation par enregistrement donneur, on suppose qu'il existe des répondants qui ont des caractéristiques semblables à celles de chaque non-répondant. Si on impute la moyenne de l'enquête courante, on suppose que la valeur moyenne pour les répondants est la même que celle qui s'applique aux non-répondants. De même, on peut étudier tous les estimateurs et dresser la liste des hypothèses implicites. Le premier estimateur proposé dans la présente sous-section vise à assouplir les hypothèses quelque peu restrictives (et habituellement fausses) imposées dans la sous-section précédente; par contre cet estimateur est plus complexe. Cet estimateur, dit "chaîn link" (estimateur en chaîne), a été proposé par Madow et Madow (1978).

Voici comment cet estimateur est obtenu. Nous supposons d'abord que le taux de variation (tendance) des populations de non-répondants et de répondants sont les mêmes que ceux qui ont été observés lors de l'enquête précédente, ce qui nous permet d'estimer la moyenne de la variable Y pour la population dans le cas de la population des non-répondants lors de l'enquête courante.

$$\bar{y}_{NRi} = \frac{\bar{y}_{NR(i-1)}}{\bar{y}_{R(i-1)}} \bar{y}_{Ri}.$$

On peut alors déterminer la valeur imputée d'après la variable auxiliaire.

$$y_{ii} = \frac{\bar{y}_{NRi}}{\bar{y}_{NRi}} x_{ii}$$

$$= \frac{\bar{y}_{NR(i-1)}}{\bar{y}_{RT}} \frac{\bar{y}_{RT}}{\bar{y}_{R(i-1)}} x_{ii}$$

Il faut remarquer que cela équivaut à une application plus complexe de la méthode du modèle de régression étudiée à la sous-section 3.3. D'abord, il faut imputer temporairement $y_{ii} = \bar{y}_{NRi}$ comme cela est précisé ci-dessus. Il faut ensuite utiliser le modèle II et définir le groupe d'imputation comme l'ensemble de tous les enregistrements des non-répondants lors de l'enquête courante pour la variable Y . La variable de réponse est Y_i . La variable explicative est X_i . L'estimateur résultant correspond au résultat de la formule donnée plus haut. Le second estimateur proposé dans la présente sous-section peut être utilisé quand on possède des données sur la variable Y pour plusieurs enquêtes antérieures. Cette méthode n'utilise pas de variables auxiliaires ou de données provenant d'autres enregistrements. On considère le comportement de chaque non-répondant indépendamment des autres. Cette

Plusieurs notes explicatives doivent accompagner la notation. D'abord, les indices R et NR sont tels que définis dans l'enquête, peu importe que chaque enregistrement corresponde à un répondant ou à un non-répondant. Deuxièmement, les valeurs des variables $y_{i(t-1)}$, $x_{i(t-1)}$ peuvent elles-mêmes avoir été imputées. La seule contrainte est qu'aucune d'entre elles ne peut être omise. Troisièmement, la notation n'inclut pas le concept de classes d'imputation. Les classes d'imputation sont essentiellement définies après les strates, en ce sens qu'elles définissent des ensembles d'enregistrements jugés homogènes à l'intérieur des groupes et hétérogènes entre les groupes. Cependant, la notation ainsi que les estimateurs d'imputation peuvent facilement être étendus ou augmentés pour inclure les classes d'imputation. Les estimateurs peuvent donc être classés selon:

- (i) le choix du modèle, I ou II,
- (ii) le groupe d'imputation,
- (iii) les variables dans la régression utilisée pour estimer le paramètre.

Les données des enregistrements qui font partie du groupe d'imputation précisé sont exactement celles utilisées pour estimer le(s) paramètre(s) dans le modèle. Ce concept permet une grande flexibilité. Par exemple, il permet d'exclure les valeurs extrêmes aberrantes du calcul de la valeur estimative du paramètre. Une fois le paramètre estimé, ce dernier est utilisé à des fins de prévision pour déterminer la valeur imputée. Dans la notation employée, y_i représente toujours la variable prévue.

D'après ces deux modèles, nous proposons huit estimateurs d'imputation et cette liste peut éventuellement être augmentée. On pourrait obtenir des estimateurs additionnels en choisissant, par exemple, d'autres modèles qui pourraient comprendre plus de variables. On peut voir, en parcourant la liste de ces estimateurs, qu'il s'agit des estimateurs d'imputation familiaires utilisés traditionnellement.

Estimateur 1: La valeur tirée de l'enquête antérieure pour la même unité est imputée.

$$y_{i(t-1)}$$
Estimateur 2: La valeur moyenne est tirée de l'enquête précédente et est imputée.

$$\bar{y}_{(t-1)}$$
Estimateur 3: La valeur moyenne, pour tous les répondants, pour l'enquête courante est imputée. \bar{y}_{iR}

Estimateur 4: La valeur est reproduite directement dans l'enregistrement receveur à partir de l'enregistrement donneur. y_D^i
Estimateur 5: Une estimation, par la méthode du quotient, qui utilise des valeurs tirées de l'enquête courante, est imputée.

$$\frac{\bar{y}_{iR}}{\bar{x}_{iR}} x_{iR}$$
Estimateur 6: Une estimation, par la méthode du quotient, basée sur les valeurs des enregistrements donneur et receveur est imputée.

$$\frac{y_D^i}{x_D^i} x_{iR}$$
Estimateur 7: La valeur tirée de l'enquête antérieure pour la même unité, avec un ajustement de la tendance calculé à partir d'une variable auxiliaire, est imputée.

Estimateur 8: La valeur tirée de l'enquête antérieure pour la même unité, avec un ajustement de la tendance calculé à partir du changement dans les valeurs déclarées pour la variable Y , est imputée.

$$\frac{\bar{y}_{iR}}{\bar{y}_{i(t-1)R}} y_{i(t-1)}$$

Le dernier point à mentionner à propos de l'imputation par enregistrement donneur est le concept d'une "pénalité" pour utilisation d'enregistrements donneurs. Cette pénalité réduirait le nombre de fois où un enregistrement donneur particulier est utilisé. Pour l'imputation par enregistrement donneur de données catégorielles, un enregistrement donneur est choisi sans remplacement, dans la population d'enregistrements donneurs. Cette stratégie doit être légèrement modifiée si la taille de la population d'enregistrements receveurs est supérieure à celle de la population d'enregistrements donneurs.

Dans le cas des données numériques, on modifie la fonction de distance en augmentant la distance selon le nombre de fois qu'un enregistrement donneur particulier est utilisé. Une façon d'atteindre ce résultat consiste à utiliser $D'(X, Y)$ pour calculer les distances, où

$$D'(X, Y) = D(X, Y) \times (1 + ud),$$

où u représente la "pénalité" imposée par l'utilisateur, d représente le nombre de fois qu'un enregistrement donneur a été choisi. L'imposition d'une pénalité pour la fonction de distance implique que le choix d'un enregistrement donneur pour chaque enregistrement receveur dépend maintenant de l'ordre des enregistrements receveurs.

3.3 Modèles de régression

La présente sous-section traite les estimateurs d'imputation obtenus à l'aide de modèles de régression. Pour cette étude, nous n'utiliserons que deux modèles. Ce sont:

MODÈLE I : $y_i = \alpha + \epsilon_i$ $\text{Var}(\epsilon_i) = \sigma^2$

MODÈLE II: $y_i = \beta x_i + \epsilon_i$ $\text{Var}(\epsilon_i) = \sigma^2 x_i$

Il faut remarquer que ces modèles constituent des cas spéciaux de la forme plus générale des modèles de régression, qui est la suivante:

$$y = \tilde{X}\tilde{\beta} + \epsilon,$$

$$\text{ou } E(\tilde{\epsilon}) = \tilde{0}, V(\tilde{\epsilon}) = \tilde{V}.$$

Le modèle II est employé quand des données auxiliaires sont disponibles. Sinon, le modèle I est utilisé. Dans les deux modèles, un paramètre doit être estimé. La méthode des moindres carrés permet d'obtenir les estimations suivantes pour les paramètres:

$$\alpha = \bar{y},$$
$$\tilde{\beta} = \frac{\bar{y}}{\bar{x}}.$$

Avant d'énoncer les divers estimateurs proposés, nous présenterons une notation particulière. Soit t l'indice inférieur pour le temps t , l'enquête actuelle; y_{it} la variable à l'étude pour l'unité i et le temps t ; c est la valeur à imputer pour les enregistrements receveurs; x_{it} la variable auxiliaire (corrélée avec Y) pour l'unité i et le temps t ; R l'indice inférieur pour tous les répondants au temps t (c à-d. que y_{it} est connu); NR l'indice inférieur pour tous les non-répondants au temps t (c à-d. que y_{it} doit faire l'objet d'une imputation), C, D les indices supérieurs qui indiquent soit un enregistrement receveur (C), soit un enregistrement donneur (D), chaque fois que la distinction est requise.

Deux types de transformations des données sont proposés ici. Dans les deux cas, chaque variable doit être transformée indépendamment. Les deux transformations proposées sont la transformée des valeurs de rang et la transformée d'échelle de localisation.

Pour la transformée des valeurs de rang, les valeurs de chaque variable sont triées, les valeurs de rang sont ensuite divisées par une constante appropriée afin que toutes les valeurs se trouvent dans l'intervalle de zéro à un. Les valeurs transformées sont réparties uniformément sur cet intervalle.

La transformée d'échelle de localisation a la forme suivante:

$$y^T = \frac{1}{b} (y - a),$$

où y^T est la valeur transformée,
 y est la valeur originale de la donnée,
 a, b sont des paramètres précisés par l'utilisateur.

Deux choix fréquents pour ces constantes sont, premièrement, que a soit la moyenne de l'échantillon et b l'écart-type de l'échantillon et, deuxièmement, que a soit la valeur minimale de l'échantillon et b l'étendue des valeurs de l'échantillon. D'autres options sont possibles.

Dans le choix d'une transformation pour les données, il faut tenir compte de la robustesse et des observations extrêmes aberrantes. La transformée des valeurs de rang est très robuste par rapport aux changements dans les valeurs des données et elle ramène les observations extrêmes aberrantes plus près des autres valeurs des données. Cela peut être ou ne pas être souhaitable. Aucune limite n'est imposée aux valeurs transformées quand on utilise la transformée d'échelle de localisation avec la moyenne et l'écart-type. Ces paramètres sont également sensibles aux valeurs extrêmes aberrantes. Le choix de la valeur minimale et l'étendue des valeurs permettrait de limiter les valeurs transformées entre zéro et un. Cependant, ces paramètres sont très sensibles aux valeurs extrêmes. Une valeur très grande pourrait donner à toutes les valeurs transformées, sauf une, une valeur très rapprochée de zéro.

Pour le choix de la fonction de distance, nous proposons une famille de fonctions de distance. Ce sont les normes \mathcal{L}^p pondérées, où p est une constante précisée par l'utilisateur. Voici la forme générale de ces fonctions:

$$D(X, Y) = \left[\sum_{k=1}^K w_k |x_k - y_k|^p \right]^{1/p},$$

où x_k, y_k sont les r variables d'appariement dans les deux enregistrements,
 w_k sont des poids précisés par l'utilisateur,
 p est une constante précisée par l'utilisateur.

Les poids sont utilisés si l'utilisateur désire que certaines des variables d'appariement contribuent plus que d'autres au calcul de la distance. Les valeurs implicites servent à donner à tous les poids la valeur un.

Trois choix particuliers d'une valeur de p présentent un intérêt spécial, $p = 1, p = 2$, et $p = \infty$. Quand $p = 1$, cette fonction calcule les distances entre les îlots. Quand $p = 2$, c'est la distance euclidienne qui est calculée. Le cas limite de cette fonction, quand $p = \infty$, donne la distance minimax. Pour ce choix de p , la fonction s'écrit de la façon suivante:

$$D(X, Y) = \max_{1 \leq k \leq r} [w_k |x_k - y_k|].$$

variables d'appariement. Dans une telle situation, il est possible (selon le soin apporté par l'utilisateur pour définir les variables d'appariement) qu'un enregistrement donneur particulier ait une variable d'appariement pour laquelle il faut effectuer une imputation. L'utilisateur qui introduit les spécifications des variables d'appariement doit savoir que cette décision peut entraîner une augmentation considérable de la charge de travail.

Nous proposons une méthode qui peut être utilisée pour déterminer automatiquement les variables d'appariement. Cette procédure peut être utilisée de façon analogue pour les données numériques ainsi que pour les données catégorielles. Voici en substance en quoi cette procédure consiste. L'ensemble des variables d'appariement doit contenir au moins les variables qui se trouvent dans les règles de contrôle avec les variables à soumettre à l'imputation. D'après la définition que nous avons déjà donnée, ce sont les contrôles actifs. Cette méthode semble intuitivement raisonnable car il est souhaitable qu'il existe une corrélation entre les variables d'appariement et les variables à soumettre à l'imputation. Les variables dans les contrôles actifs limitent la gamme de valeurs possibles à imputer. Cela implique un certain type de dépendance ou de structure de corrélation.

L'emploi de cette procédure d'appariement avec la transcription directe a une conséquence importante pour les variables catégorielles. Nous sommes assurés que toutes les variables imputées seront acceptées aux contrôles. Cela est très important car c'est une condition requise pour créer un enregistrement épuré. Sans cette assurance, l'utilisateur doit effectuer un nouveau contrôle des enregistrements et peut-être adopter une procédure d'imputation secondaire. Pour les données numériques, la similarité, telle que définie par une fonction de distance, ne garantit pas ce résultat. Cependant, plus la distance entre les enregistrements donneur et receveur approche de zéro, plus la probabilité que les valeurs imputées seront acceptées aux contrôles est élevée.

Un exemple nous permettra d'illustrer la détermination des variables d'appariement au moyen de cette procédure automatisée.

Supposons les cinq contrôles suivants:

$$I. \quad A + B \leq \alpha_1,$$

$$II. \quad B - E \leq \alpha_2$$

$$III. \quad C + 2D + 3E \leq \alpha_3,$$

$$IV. \quad A + C + D \leq \alpha_4,$$

$$V. \quad A - 2B + C \leq \alpha_5.$$

Il y a cinq variables d'enquête A, B, C, D, E et $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ sont des scalaires connus.

Pour l'enregistrement receveur à l'étude, il faut faire une imputation seulement pour la variable B .

La première étape consiste à identifier les contrôles actifs. Dans cet exemple, il y en a trois. Ce sont les contrôles I, II et V.

La deuxième étape consiste à déterminer les variables actives. On entend par variable active toute variable contenue dans au moins un des contrôles actifs. Dans l'exemple, il y a quatre variables actives: A, B, C, E . Il faut remarquer que, par définition, toutes les variables à soumettre à l'imputation se trouvent parmi les variables actives.

La troisième étape consiste à déterminer les variables d'appariement; ce sont les variables actives pour lesquelles aucune imputation n'est requise. Dans cet exemple, les variables d'appariement sont A, C, E .

L'imputation par enregistrement donneur pour les données numériques exige, en plus de la détermination des variables d'appariement, le choix d'une transformation des données ainsi que d'une fonction de distance.

3.2 Imputation par enregistrement donneur

L'imputation par enregistrement donneur est une méthode qui apparie chaque enregistré-ment pour lequel on doit effectuer une imputation, l'enregistré-ment receveur, avec un enregistré-ment d'une population d'enregistré-ments donneurs définie. Une des méthodes utilisées pour déterminer la valeur à imputer consiste à reproduire directement dans l'enregistrement receveur la valeur tirée de l'enregistrement donneur. Pour les variables numériques, si les renseignements auxiliaires appropriés sont disponibles, on peut utiliser des méthodes plus complexes pour déterminer la valeur à imputer. Les estimateurs pour l'imputation par enregistrement donneur sont étudiés plus en détail dans la sous-section 3.3. Habituellement, la population des enregistré-ments donneurs est définie comme l'ensemble de tous les enregistré-ments de l'enquête en cours pour lesquels aucune variable ne doit être soumise à l'imputation. Nous reportant au cadre de prévisions décrit au début de la section 3, cette situation implique que $f(X_1, \dots, X_N)$ est la fonction de probabilité empirique. Cependant, il est possible d'utiliser d'autres méthodes pour définir la population d'enregistré-ments donneurs. Pour la suite de la discussion sur l'imputation par enregistrement donneur, nous supposons tout simplement qu'une population d'enregistré-ments donneurs a été définie.

Les paires d'enregistrement donneur – enregistré-ment receveur sont formées au moyen de variables d'appariement. Les variables d'appariement sont définies comme des variables pour lesquelles aucune imputation ne doit être effectuée pour l'enregistrement receveur et pour lesquelles il existe une corrélation élevée avec la ou les variables à soumettre à l'imputation. De préférence, la corrélation entre les variables d'appariement doit en outre être faible. Deux variables d'appariement fortement corrélées auraient le même pouvoir discriminatif qu'une seule, mais elles auraient pour effet de doubler le poids donné à l'une quelconque de ces variables.

Dans le cas des variables catégorielles, un enregistré-ment donneur est choisi, au moyen d'une méthode aléatoire quelconque, parmi les enregistré-ments donneurs potentiels qui ont, pour les variables d'appariement, les mêmes valeurs que celles de l'enregistrement receveur. Comme les variables numériques peuvent prendre un nombre beaucoup plus considérable de valeurs que les variables catégorielles, il est très peu probable qu'il existera un appariement exact pour les variables d'appariement. Par conséquent, pour les données numériques, on utilise une fonction de distance pour définir la similarité. Cette fonction de distance est une fonction des variables d'appariement des enregistré-ments receveurs et des enregistré-ments donneurs potentiels. L'enregistrement donneur choisi est celui qui se trouve à la plus faible distance de l'enregistrement receveur. Habituellement, les variables d'appariement sont transformées pour le calcul des distances afin de supprimer l'effet de l'échelle dans laquelle la variable est mesurée. Par exemple, ce serait très ennuyeux pour l'utilisateur si la formation du couple enregistré-ment donneur – enregistré-ment receveur dépendait du fait que la variable de longueur est mesurée en mètres ou en pieds. Les transformations et fonctions de distance proposées sont étudiées plus loin.

Les variables d'appariement à utiliser peuvent être définies par l'utilisateur ou déterminées au moyen d'une procédure automatisée. Habituellement, à cause de contraintes de temps, toutes les décisions doivent être prises avant la collecte des données. Par conséquent, si la détermination des variables d'appariement est effectuée par l'utilisateur, ce dernier doit définir les variables d'appariement pour chaque combinaison possible de variables à soumettre à l'imputation. Si le fichier contient N variables, l'utilisateur doit définir $(2^N - 2)$ spécifications. Il est évident que la valeur N n'a pas à être très grande pour que cette méthode devienne insupportable à appliquer. Pour réduire ce nombre, il est possible de définir des variables d'appariement par strates. Tous les enregistré-ments receveurs d'une strate particulière utiliseraient les mêmes

les données. Les contrôles actifs sont définis comme étant le sous-ensemble de contrôles dans lequel figurent les variables à soumettre à l'imputation. Cela peut aussi être exprimé dans la notation du cadre de prévisions donné au début de la section 3. La distribution conditionnelle $f(X_1, \dots, X_{m_l} | X_{m_l+1}, \dots, X_N)$ déterminera une valeur unique pour certaines ou pour toutes les variables X_1, \dots, X_{m_l} .

Donnons un exemple pour illustrer la procédure utilisée pour identifier l'imputation déterministe. Il faut remarquer que bien que l'exemple utilise des variables numériques, une situation analogue existe pour les variables catégorielles.

Considérons les trois contrôles suivants:

$$X + Y \leq 16,$$

$$Y + Z \leq 4,$$

$$X - 3Z \leq 8.$$

L'enregistrement considéré possède les valeurs suivantes:

$$X = 11 \text{ et } Y = 3.$$

Il faut faire une imputation pour la variable Z.

Il n'est pas évident si l'imputation déterministe peut être utilisée. Dans une première étape, on doit considérer tous les contrôles actifs. Dans l'exemple, deux contrôles contiennent la variable Z.

$$Y + Z \leq 4,$$

$$X - 3Z \leq 8.$$

Il faut ensuite introduire les valeurs connues de X et Y dans ces contrôles pour déterminer la région d'acceptation réduite.

$$3 + Z \leq 4,$$

$$11 - 3Z \leq 8.$$

En réduisant ces inégalités nous obtenons la solution suivante.

$$Z \leq 1,$$

$$Z \geq 1.$$

Il est maintenant évident que $Z = 1$ est la seule valeur imputée valide qui soit possible. Dans la plupart des situations réelles, l'incidence de l'imputation déterministe devrait être faible. Le contraire indiquerait que les contrôles sont plus restrictifs qu'il n'est nécessaire ou souhaitable, et cela devrait entraîner un réexamen des spécifications de contrôle. Cependant, comme l'imputation déterministe permet de réduire le problème de l'imputation, elle constitue une première étape nécessaire.

3. TECHNIQUES D'IMPUTATION PROPOSÉES

La présente section comprend quatre sous-sections qui définissent toutes les techniques d'imputation proposées. Ce sont l'imputation déterministe, l'imputation par enregistrement donneur, les modèles de régression et les autres estimateurs pour l'imputation. L'utilisation des modèles de régression ainsi que la sous-section sur les autres estimateurs ne s'applique qu'aux données numériques. Les deux autres sous-sections s'appliquent aux données numériques et aux données catégorielles.

Presque toutes les techniques d'imputation peuvent être formulées selon un cadre de prévisions décrit de la façon suivante par Rubin (1976). On précise une distribution à plusieurs variables, $f(X_1, \dots, X_N)$, qui résume le comportement statistique de la population des enregistrements complets. Cela peut être fait, que les variables individuelles soient quantitatives ou qualitatives. Sans perte de généralité, pour une enregistrement i pour lequel il faut effectuer une imputation, les N variables peuvent être divisées de la façon suivante: X_1, \dots, X_{m_i} sont les variables pour lesquelles il faut effectuer une imputation et X_{m_i+1}, \dots, X_N , celles pour lesquelles aucune imputation n'est nécessaire. On peut alors obtenir une distribution conditionnelle $f(X_1, \dots, X_{m_i} \mid X_{m_i+1}, \dots, X_N)$. Les valeurs imputées, y_1, \dots, y_{m_i} , sont choisies pour X_1, \dots, X_{m_i} à partir de l'ensemble

$$\{y_1, \dots, y_{m_i} : f(y_1, \dots, y_{m_i} \mid x_{m_i+1}, \dots, x_N) > 0\}$$

On peut employer divers mécanismes de sélection. Cependant, comme nous l'avons mentionné ci-dessus, quelques-uns ne s'appliquent qu'à certains types de données.

Il faut remarquer que ces propositions ne contiennent rien de neuf ou de radicalement différent. Elles sont basées sur des travaux effectués auparavant à Statistique Canada et à l'extérieur. La discussion de l'imputation par enregistrement donneur est basée sur l'article de Fellegi et Holt (1976). La façon de déterminer une valeur à imputer au moyen d'un modèle est étudiée par Little (1982). D'autres articles connexes qui présentent un certain intérêt sont ceux de Sande (1976), de Kalton et Kasprzyk (1982) et de Kalton et Kish (1981).

3.1 Imputation déterministe

Le premier type d'imputation est appelé imputation déterministe. C'est le cas où une seule valeur peut être acceptée au contrôle. Si plus d'une variable doit faire l'objet d'une imputation dans un enregistrement particulier, une solution déterministe peut être possible pour certaines variables ou pour toutes. On doit vérifier si l'imputation déterministe peut être utilisée avant de passer à d'autres procédures d'imputation.

L'imputation déterministe peut se présenter dans des situations très simples et faciles à déterminer. Supposons, par exemple, qu'il existe un contrôle $A + B = 10$. Pour l'enregistrement étudié, A doit recevoir une valeur imputée et B a la valeur 6. Il est évident que $A = 4$ est la seule valeur qui puisse être acceptée au contrôle. Un autre exemple illustre les cas de ce genre pour les variables catégorielles. Supposons qu'un contrôle est énoncé de la façon suivante: "Si le lien avec la personne repère du ménage est épouse, alors le sexe doit être féminin". Si l'enregistrement repère contient "épouse" comme valeur pour le "lien avec la personne repère du ménage" et s'il faut faire une imputation pour la variable "Sexe", la seule valeur imputée valide est Sexe = Féminin.

Cependant, dans une enquête typique, on trouvera plusieurs contrôles plutôt qu'un seul. Cela peut signifier qu'une solution déterministe existante peut ne pas être apparente. La procédure utilisée pour vérifier si une imputation déterministe est possible consiste à trouver la région d'acceptation réduite définie par les contrôles actifs et les "bonnes" valeurs pour

autres, il peut orienter son choix par des données chronologiques, en testant les diverses solutions avant de recueillir les données. Cette opération n'entraîne pas d'effort superflu. Une fois que le système généralisé est sur pied, les utilisateurs n'ont plus à mobiliser autant de ressources et le délai d'application s'en trouve raccourci.

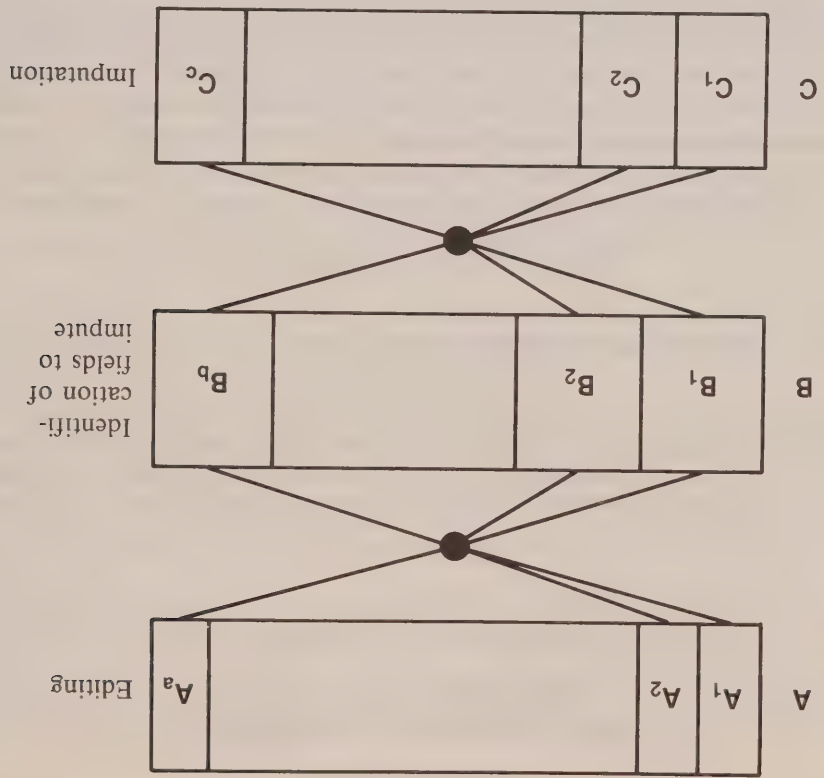
L'application d'un système généralisé à toutefois ses inconvénients. Dans un contexte de production, un logiciel général pourra s'avérer moins efficace que le système sur mesure qui aurait autrement été utilisé. Au début, un système généralisé nécessitera plus de ressources qu'un système personnalisé. Mais le coût plus élevé doit être considéré à la lumière du fait qu'il est encore plus cher de créer constamment des programmes sur mesure. Par ailleurs, il n'est pas raisonnable de penser qu'un système généralisé puisse répondre à toutes les conditions. Deux possibilités s'offrent alors à l'utilisateur. D'une part, il peut concevoir son propre module. Cette solution n'exige pas autant d'effort qu'une personnalisation complète. Toutefois, si cela se produit fréquemment, le système généralisé n'a plus sa raison d'être. D'autre part, l'utilisateur peut adapter ses exigences au système généralisé. Si le système a été bien conçu, une telle adaptation ne devrait pas avoir d'effets notables sur la qualité des données. Il convient aussi de souligner qu'il est souvent nécessaire de modifier les exigences originales dans l'élaboration d'un système personnalisé.

2. DONNÉES DE BASE SUR L'IMPUTATION

Dans le présent document, le mot "imputation" se rapporte à une certaine catégorie de procédures utilisées pour traiter les cas de non-réponse. Les données en entrée sont constituées d'un fichier de données saisies. La procédure d'imputation crée un fichier dont chaque enregistrement est "épuré"; un enregistrement "épuré" est un enregistrement où aucune donnée ne manque et qui est acceptée par tous les contrôles précisés. Si l'on veut créer un enregistrement épuré, une valeur doit être estimée pour toute valeur manquante.

Les contrôles précisés par l'utilisateur sont des contraintes logiques imposées aux valeurs que chaque variable peut prendre. Tous les contrôles, ensemble, définissent la région d'acceptation des données. Pour les données catégorielles, un contrôle est précisé comme un ensemble de combinaisons de valeurs de données acceptables. La région d'acceptation peut être représentée comme un ensemble de sommets dans un espace à N dimensions. Pour les données numériques, un contrôle est une égalité ou une inégalité linéaire. Le fait d'imposer la linéarité ne constitue pas une exigence trop restrictive, car un contrôle non linéaire peut être rendu linéaire au moyen de manipulations algébriques ou par l'addition de variables supplémentaires qui sont des fonctions non linéaires, définies de façon appropriée, des variables d'enquête. La région d'acceptation, dans le cas des données numériques, est un ensemble de régions convexes dans un espace à N dimensions. Il peut y avoir plus d'une région convexe parce qu'il est possible qu'il existe des contrôles conditionnels. Ces contrôles ne se rapportent qu'à un sous-ensemble d'enregistrements; par exemple les contrôles pertinents à un enregistrement particulier peuvent être fort différents selon que la valeur "Masculin" ou "Féminin" est donnée à la variable "Sexe". Si un enregistrement particulier est rejeté à un ou plusieurs contrôles, il se peut qu'on ne puisse déterminer quelles variables sont erronées et doivent, par conséquent, faire l'objet d'une imputation. Par exemple, la combinaison $A + B \leq C$ sera rejetée au contrôle si l'enregistrement considère possède les valeurs $A = 10$, $B = 5$, $C = 12$. Sept combinaisons de variables peuvent être modifiées pour obtenir un enregistrement épuré. Ce sont A , B , C , $A \& B$, $A \& C$, $B \& C$, et $A \& B \& C$. Sans autre renseignement ou règle de décision, chacun de ces choix est également valide. Nous n'étudierons pas dans le présent document le problème de la détermination des variables à soumettre à l'imputation. Nous supposons que, pour chaque enregistrement, ces variables ont été identifiées. Aucune distinction n'est faite ici entre les variables soumises à l'imputation à cause de valeurs manquantes et celles qui le sont à cause de rejets au contrôle.

Figure 1. Exemple de système généralisé - Contrôle et imputation



Elaborer un système de traitement par une méthode modulaire a des conséquences importantes. On peut dire que le système est en quelque sorte toujours en développement puisqu'on peut et qu'en principe on devrait sans cesse y incorporer de nouveaux modules qui renferment des méthodes inédites et apportent des perfectionnements aux anciens modules. Cette extensibilité fait que le système se prête bien à la création de prototypes, notion très importante dans le domaine. La création de prototypes est une méthode par laquelle on crée un premier sous-ensemble de modules. Le système devient alors accessible à un certain nombre d'utilisateurs. D'autres modules viennent s'ajouter par la suite afin de répondre aux besoins d'autres utilisateurs. Ainsi, le principal avantage de la modularité et de la création de prototypes est qu'elles permettent d'envisager des améliorations graduelles du système et facilitent ces améliorations. Il y a toutefois une condition minimale mais indispensable à l'application de telles méthodes. Il est en effet nécessaire de définir soigneusement un système (comme celui représenté à la figure 1) et un cadre d'utilisation central (structure des fichiers de données et langage de programmation) et de les définir au tout début du processus général d'élaboration. Outre les avantages liés à l'élaboration du système, il y a ceux qui peuvent découler de sa mise en application. Les possibilités qui s'offrent à l'utilisateur sont nombreuses. Si celui-ci doit choisir parmi plusieurs solutions qui semblent toutes aussi viables les unes que les

Méthodes d'imputation dans un système généralisé

P. GILES et C. PATRICK¹

RÉSUMÉ

Afin de répondre aux exigences de traitement de la plupart de ses enquêtes, Statistique Canada a mis sur pied un projet visant à élaborer un système généralisé de contrôle et d'imputation. Les auteurs de ce document analysent les diverses méthodes d'imputation qui ont été proposées dans le cadre de ce projet, pour traiter la non-réponse partielle. Ils se penchent aussi sur les aspects importants de l'application de ces propositions dans un système généralisé.

MOTS CLÉS: Modularité; imputation par enregistrement; prototype; modèles de régression.

1. SYSTEMES GENERALISES

Comme les ressources consacrées aux enquêtes sont de plus en plus limitées depuis quelques années, particulièrement en ce qui a trait à la recherche, on parle de plus en plus maintenant de logiciel général. Un logiciel général est une série de programmes machines intégrés en un système et qui permet à l'utilisateur de résoudre son problème par la solution la plus appropriée. Par exemple, un utilisateur doit prélever un échantillon d'enregistrements dans un fichier de données. Un système généralisé d'échantillonnage offrirait à l'utilisateur divers plans de sondage, par exemple l'échantillonnage aléatoire simple ou l'échantillonnage avec probabilités inégales (avec ou sans remise), l'échantillonnage systématique, l'échantillonnage stratifié ou encore le sondage en grappes.

Un système véritablement généralisé est presque par définition une chose complexe. La modularité est un moyen essentiel de réduire la complexité du système en divisant la tâche globale en un certain nombre de sous-tâches moins complexes. Chacune des sous-tâches ou fonctions est exécutée séquentiellement. De plus, à l'intérieur de chaque sous-tâche, l'utilisateur peut choisir parmi diverses méthodes d'exécution celle qui lui convient le mieux. Ainsi, non seulement est-il possible de diviser la tâche globale en tâches plus simples, mais il existe aussi plusieurs façons d'exécuter chaque sous-tâche.

La figure 1 montre comment le contrôle et l'imputation peuvent être répartis en trois sous-tâches. Ces trois sous-tâches sont le contrôle, la définition des zones à imputer et l'imputation proprement dite. Chaque section ou module d'une sous-tâche correspond à une méthode d'exécution particulière. Par exemple, C1 pourrait correspondre à une méthode quelconque d'imputation par enregistrements donneurs, C2 pourrait correspondre à une méthode d'imputation par la moyenne, et ainsi de suite. L'utilisateur choisirait un module par sous-tâche (A, B et C). Il convient de souligner que le système représenté ci-dessus n'est pas le seul système de contrôle et d'imputation qui puisse exister. En fait, le système proposé dans le projet de Statistique Canada comporte cinq sous-tâches plutôt que trois. Si nous avons choisi d'illustrer un système à trois sous-tâches, ce n'est que pour des raisons de simplicité.

Chaque sous-tâche illustrée à la figure 1 représente une fonction très précise. Les fichiers d'entrée nécessaires et les fichiers résultant doivent avoir une structure préalable. L'utilisateur peut ainsi concentrer son attention sur le choix d'un module dans chaque sous-tâche, sachant

¹ Philip Giles et Charles Patrick, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Parc Tunney, Ottawa (Ontario), Canada, KIA 0T6.

BIBLIOGRAPHIE

- HEITJAN, D.F., et RUBIN, D.B. (1986). Inference for coarse data using multiple imputation. Dans *Proceedings of the 18th Symposium on the Interface of Computer Science and Statistics*. Herzog, T.N., et RUBIN, D.B. (1983). Using multiple imputations to handle nonresponse in sample surveys. Dans *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliography*, New York: Academic Press, 209-245.
- LI, K.H. (1985). *Hypothesis Testing in Multiple Imputation - with Emphasis on Mixed-up Frequencies in Contingency Tables*. Thèse de doctorat, Département de statistique, Université de Chicago.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- RUBIN, D.B. (1977). The design of a general and flexible system for handling nonresponse in sample surveys. Document non publié à l'usage du U.S. Social Security Administration.
- RUBIN, D.B. (1978). Multiple imputations in sample surveys — a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20-34 et dans *Imputation and Editing of Faulty or Missing Survey Data*, U.S. Dept. of Commerce, 1-23.
- RUBIN, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. *Proceedings of the 1979 Meetings of the ISI-IASS, Manille*.
- RUBIN, D.B. (1980). *Handling Nonresponse in Sample Surveys by Multiple Imputations*. U.S. Dept. of Commerce, Bureau of the Census Monograph.
- RUBIN, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.
- RUBIN, D.B. (1986a). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- RUBIN, D.B. (1986b). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 87-94.
- RUBIN, D.B., et SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- SCHENKER, N. (1985). *Multiple Imputation for Interval Estimation from Surveys with Ignorable Nonresponse*. Thèse de doctorat, Département de statistique, Université de Chicago.

REMERCIEMENTS

Ce travail a été subventionné par le National Science Foundation, Subvention SES-8311428. L'auteur tient à remercier M.P. Singh et un arbitre pour leurs commentaires lors de la rédaction de cet article.

Tableau 10

Valeur (en pourcentage) de D_M fondée sur une distribution de $F_{k,v}$, (distribution de référence), en fonction du niveau nominal (α), du nombre d'éléments testés (k), du nombre d'imputations itératives propres (M) et de la proportion de données manquantes (γ).

k	M	$\gamma =$	$\alpha = 1\%$						$\alpha = 5\%$						$\alpha = 10\%$								
			2	3	5	10	25	50	100	2	3	5	10	25	50	100	2	3	5	10	25	50	100
2	2	1.0	1.2	1.6	2.5	4.9	5.3	5.9	7.5	9.9	10.3	12.9	10.9	11.0	10.4	9.9	9.9	10.0	10.2	10.4	11.0	10.0	10.0
	3	1.0	1.0	1.0	1.3	4.9	4.9	5.0	5.5	9.9	9.8	10.0	10.9	10.9	10.2	9.8	9.9	10.0	10.0	10.4	11.0	10.0	10.0
	5	1.0	1.0	1.1	1.2	5.0	5.1	5.6	6.2	10.1	10.0	10.8	12.2	11.6	10.6	10.3	10.1	10.0	10.0	10.4	11.0	10.0	10.0
	10	1.0	1.0	1.1	1.2	5.0	5.2	5.3	5.9	10.1	10.3	10.6	11.6	11.6	10.6	10.3	10.1	10.0	10.0	10.4	11.0	10.0	10.0
	25	1.0	1.0	1.1	1.2	5.0	5.1	5.2	5.6	10.1	10.2	10.4	10.9	10.9	10.4	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
	50	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.1	10.0	10.0	10.0	10.2	10.2	10.0	10.0	10.0	10.0	10.0	10.2	10.4	11.0	10.0
3	2	1.0	1.1	1.3	1.7	5.1	5.3	5.6	6.3	10.3	10.6	12.0	12.3	11.1	10.6	10.3	10.0	10.0	10.0	10.4	11.0	10.0	10.0
	3	1.0	1.0	1.0	1.0	5.1	5.2	5.3	5.7	10.2	10.5	10.9	12.3	12.2	10.9	10.3	10.1	10.0	10.0	10.4	11.0	10.0	10.0
	5	1.0	1.0	1.1	1.3	5.0	5.2	5.4	6.2	10.1	10.3	10.8	12.2	11.6	10.6	10.3	10.1	10.0	10.0	10.4	11.0	10.0	10.0
	10	1.0	1.0	1.1	1.2	5.0	5.2	5.3	5.9	10.1	10.3	10.6	11.6	11.6	10.6	10.3	10.1	10.0	10.0	10.4	11.0	10.0	10.0
	25	1.0	1.0	1.1	1.2	5.0	5.1	5.2	5.6	10.1	10.2	10.4	10.9	10.9	10.4	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
	50	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.1	10.0	10.0	10.0	10.2	10.2	10.0	10.0	10.0	10.0	10.0	10.2	10.4	11.0	10.0
5	2	0.9	0.8	0.8	0.9	5.1	4.8	4.5	4.0	10.5	10.4	9.2	14.4	14.4	12.1	11.3	10.5	10.0	10.0	10.4	11.0	10.0	10.0
	3	1.0	1.0	1.0	0.9	5.2	5.5	5.7	6.1	10.5	11.3	12.1	15.4	15.4	12.2	11.1	10.4	10.0	10.0	10.4	11.0	10.0	10.0
	5	1.1	1.1	1.2	1.4	5.2	5.6	6.1	7.7	10.4	11.1	12.2	15.4	14.4	12.2	11.1	10.4	10.0	10.0	10.4	11.0	10.0	10.0
	10	1.0	1.1	1.2	1.5	5.1	5.3	5.6	6.9	10.1	10.4	11.1	13.1	13.1	10.6	10.3	10.1	10.0	10.0	10.4	11.0	10.0	10.0
	25	1.0	1.0	1.1	1.3	5.0	5.2	5.3	6.0	10.1	10.3	10.6	11.5	11.5	10.6	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
	50	1.0	1.0	1.0	1.1	5.0	5.1	5.1	5.4	10.0	10.1	10.2	10.7	10.7	10.2	10.0	10.0	10.0	10.0	10.2	10.4	11.0	10.0
10	2	0.8	0.5	0.3	0.1	5.1	4.0	2.9	1.5	10.8	10.1	5.4	16.2	16.2	13.8	12.7	11.3	10.0	10.0	10.4	11.0	10.0	10.0
	3	1.1	0.9	0.6	0.3	5.6	5.9	5.7	4.9	11.3	12.7	13.8	16.2	16.2	13.8	12.7	11.3	10.0	10.0	10.4	11.0	10.0	10.0
	5	1.1	1.2	1.3	1.4	5.4	6.3	7.4	11.0	10.7	12.4	14.8	22.7	19.0	13.4	11.0	10.0	10.0	10.2	10.4	11.0	10.0	10.0
	10	1.1	1.2	1.4	2.2	5.2	5.8	6.8	10.3	10.4	11.4	13.1	19.0	13.4	11.0	10.0	10.0	10.0	10.2	10.4	11.0	10.0	10.0
	25	1.0	1.1	1.2	1.6	5.0	5.2	5.6	7.1	10.0	10.4	11.0	13.4	13.4	10.6	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
	50	1.0	1.0	1.1	1.3	5.0	5.1	5.4	6.1	10.0	10.2	10.6	11.8	11.8	10.6	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
25	2	0.8	0.5	0.3	0.1	5.1	4.0	2.9	1.5	10.8	10.1	5.4	16.2	16.2	13.8	12.7	11.3	10.0	10.0	10.4	11.0	10.0	10.0
	3	1.1	0.9	0.6	0.3	5.6	5.9	5.7	4.9	11.3	12.7	13.8	16.2	16.2	13.8	12.7	11.3	10.0	10.0	10.4	11.0	10.0	10.0
	5	1.1	1.2	1.3	1.4	5.4	6.3	7.4	11.0	10.7	12.4	14.8	22.7	19.0	13.4	11.0	10.0	10.0	10.2	10.4	11.0	10.0	10.0
	10	1.1	1.2	1.4	2.2	5.2	5.8	6.8	10.3	10.4	11.4	13.1	19.0	13.4	11.0	10.0	10.0	10.0	10.2	10.4	11.0	10.0	10.0
	25	1.0	1.1	1.2	1.6	5.0	5.2	5.6	7.1	10.0	10.4	11.0	13.4	13.4	10.6	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
	50	1.0	1.0	1.1	1.3	5.0	5.1	5.4	6.1	10.0	10.2	10.6	11.8	11.8	10.6	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
50	2	0.8	0.5	0.3	0.1	5.1	4.0	2.9	1.5	10.8	10.1	5.4	16.2	16.2	13.8	12.7	11.3	10.0	10.0	10.4	11.0	10.0	10.0
	3	1.1	0.9	0.6	0.3	5.6	5.9	5.7	4.9	11.3	12.7	13.8	16.2	16.2	13.8	12.7	11.3	10.0	10.0	10.4	11.0	10.0	10.0
	5	1.1	1.2	1.3	1.4	5.4	6.3	7.4	11.0	10.7	12.4	14.8	22.7	19.0	13.4	11.0	10.0	10.0	10.2	10.4	11.0	10.0	10.0
	10	1.1	1.2	1.4	2.2	5.2	5.8	6.8	10.3	10.4	11.4	13.1	19.0	13.4	11.0	10.0	10.0	10.0	10.2	10.4	11.0	10.0	10.0
	25	1.0	1.1	1.2	1.6	5.0	5.2	5.6	7.1	10.0	10.4	11.0	13.4	13.4	10.6	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
	50	1.0	1.0	1.1	1.3	5.0	5.1	5.4	6.1	10.0	10.2	10.6	11.8	11.8	10.6	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
100	2	0.8	0.5	0.3	0.1	5.1	4.0	2.9	1.5	10.8	10.1	5.4	16.2	16.2	13.8	12.7	11.3	10.0	10.0	10.4	11.0	10.0	10.0
	3	1.1	0.9	0.6	0.3	5.6	5.9	5.7	4.9	11.3	12.7	13.8	16.2	16.2	13.8	12.7	11.3	10.0	10.0	10.4	11.0	10.0	10.0
	5	1.1	1.2	1.3	1.4	5.4	6.3	7.4	11.0	10.7	12.4	14.8	22.7	19.0	13.4	11.0	10.0	10.0	10.2	10.4	11.0	10.0	10.0
	10	1.1	1.2	1.4	2.2	5.2	5.8	6.8	10.3	10.4	11.4	13.1	19.0	13.4	11.0	10.0	10.0	10.0	10.2	10.4	11.0	10.0	10.0
	25	1.0	1.1	1.2	1.6	5.0	5.2	5.6	7.1	10.0	10.4	11.0	13.4	13.4	10.6	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0
	50	1.0	1.0	1.1	1.3	5.0	5.1	5.4	6.1	10.0	10.2	10.6	11.8	11.8	10.6	10.3	10.1	10.0	10.0	10.2	10.4	11.0	10.0

En définitive, l'imputation multiple est une méthode inédite de traitement de la non-réponse qui donne des signes très prometteurs. Bien qu'il y ait encore beaucoup à faire avant que cette méthode soit reçue de tous, de nombreuses constatations théoriques et pratiques intéressantes donnent à penser que les efforts consacrés à son perfectionnement trouveront leur récompense dans l'avancement de la statistique appliquée.

distribution normale et la variance inter-imputation est nulle.

$1 - \alpha$	M	1	2	3	5	∞	1	2	3	5	∞	1	2	3	5	∞	1	2	3	5	∞	1	2	3	5	∞					
.9	.1	46	42	38	34	30	26	22	18	.7	.8	.9	46	42	38	34	30	26	22	18	.7	.8	.9	46	42	38	34	30	26	22	18
	.2	42	38	34	30	26	22	18	12				42	38	34	30	26	22	12	50				42	38	34	30	26	22	12	50
	.3	38	34	30	26	22	18	12	50				38	34	30	26	22	18	50	50				38	34	30	26	22	18	50	50
	.4	34	30	26	22	18	12	50	50				34	30	26	22	18	50	50	50				34	30	26	22	18	50	50	50
	.5	30	26	22	18	12	50	50	50				30	26	22	18	12	50	50	50				30	26	22	18	50	50	50	50
	.6	26	22	18	12	50	50	50	50				26	22	18	12	50	50	50	50				26	22	18	50	50	50	50	50
	.7	22	18	12	50	50	50	50	50				22	18	12	50	50	50	50	50				22	18	12	50	50	50	50	50
	.8	18	12	50	50	50	50	50	50				18	12	50	50	50	50	50	50				18	12	50	50	50	50	50	50
	.9	12	50	50	50	50	50	50	50				12	50	50	50	50	50	50	50				12	50	50	50	50	50	50	50

matrices des variances-covariances, U_i , et en appliquant les équations précédentes sous une forme multidimensionnelle.

À cet égard, Li (1985) et Rubin (1986a) proposent une méthode simple et efficace dans le cas où M est beaucoup plus grand que k . Cette méthode consiste à définir la probabilité d'une valeur de Θ moins probable que la valeur observée, étant donné l'hypothèse nulle, comme $\text{Prob} \{F_{k,v} \mid > D_M F_{k,v}\} > \Theta_0 - \Theta_M)^T \Theta_0$ est la valeur de Θ sous l'hypothèse nulle et V est défini en utilisant l'élément diagonal moyen de $\text{with } B_M U_M^{-1}$, c'est-à-dire $\text{trace} (B_M U_M^{-1}) / k$. Des méthodes plus efficaces sont décrites dans Rubin (1986a). Des valeurs de p moins précises peuvent être obtenues directement des M seuils de signification pour données complètes découlant d'une imputation multiple; à ce sujet, voir aussi Rubin (1986a).

4. ANALYSE

4.1 Évaluation de fréquence

Bien que les inférences découlant de l'imputation multiple trouvent surtout leur justification dans une perspective bayésienne, il est possible de démontrer qu'elles ont des caractéristiques de fréquence satisfaisantes. De fait, la méthode d'imputation propre implique par définition que les inférences découlant de l'imputation multiple lorsque M est défini seront valables pour de grands échantillons. Toutefois, comme les facteurs de correction pour un nombre fini d'imputations M sont déterminés au moyen d'approximations de distributions bayésiennes a posteriori, il peut en l'occurrence se produire des imperfections. Par exemple, l'efficacité de Θ_M par rapport à celle de Θ_∞ pour les grands échantillons, c'est-à-dire le rapport entre l'efficacité de l'estimateur pour un nombre fini d'imputation M obtenues par des méthodes d'imputation propres et l'efficacité de l'estimateur pour un nombre infini d'imputation, est définie (en unités d'erreur type) comme $(1 + \gamma/M)^{-1/2}$. Même pour des valeurs de γ relativement élevées, de faibles valeurs de M produisent des estimations Θ_M presque parfaitement efficaces.

4.2 Couverture des intervalles de confiance

Pour les grands échantillons, le taux de couverture des intervalles de confiance des méthodes d'imputation propres fondés sur la distribution de t (distribution de référence) peut être établi en fonction de M , de γ et du niveau nominal $1 - \alpha$. Le tableau 9 est tiré du Rubin (1986a); il est aussi reproduit partiellement dans Rubin et Schenker (1986) et Schenker (1985). Ce tableau contient également des données pour l'imputation simple, auquel cas la variance inter-imputation est définie comme nulle puisqu'elle ne peut être estimée et la distribution de référence est une distribution normale puisque l'on ne peut estimer v sans connaître B_M . Même dans des cas extrêmes, deux ou trois imputations itératives suffisent à produire des intervalles de confiance relativement acceptables; cela tranche nettement avec l'imputation simple. Dans ce dernier cas, l'application de meilleurs méthodes de prévision, comme l'imputation par la moyenne, aurait produit des taux de couverture encore moindres.

4.3 Niveaux de signification

La recherche visant à définir avec précision des niveaux de signification est encore jeune. Le tableau 10 est tiré de Rubin (1986a); il est aussi reproduit partiellement dans Li (1985). Ce tableau indique que l'on peut réaliser des tests précis à l'aide de la fonction D_M si $M > k$ et γ est peu élevé. Des meilleurs méthodes sont examinées dans Li (1985) et Rubin (1986a) ainsi que dans une thèse que rédige actuellement T.E. Raghunathan.

3.3 Analyse – Inférence découlant des imputations itératives

Les méthodes générales servant à analyser une séries de données obtenue par imputation multiple supposent implicitement l'utilisation de méthodes d'imputation propres. Comme nous l'avons vu à la section 2, les imputations itératives obtenues dans chaque modèle sont analysées en bloc pour produire une seule inférence. Chaque série de données complètes obtenue par imputation est soumise à la méthode d'analyse de données complètes qui aurait été utilisée s'il n'y avait pas eu de non-réponse. De façon plus précise, supposons que $\theta_i, U_i, (i = 1, \dots, M)$ désignent respectivement M estimations de données complètes et les variances correspondantes pour un paramètre θ , ces valeurs étant calculées à l'aide des M séries de données obtenues par imputation multiple suivant un modèle de non-réponse. L'estimation finale de θ est

$$\bar{\theta}_M = \sum_M^{\ell=1} \theta_i / M.$$

La variance de cette estimation a deux composantes: la variance intra-imputation moyenne,

$$\bar{U}_M = \sum_M^{\ell=1} U_i / M,$$

et la variance inter-imputation,

$$B_M = \Sigma (\theta_i - \bar{\theta}_M)^2 / (M-1)$$

où $(\bullet)^2$ est remplacé par $(\bullet)^T(\bullet)$ lorsque θ est un vecteur. La variance totale de l'estimation $\bar{\theta}_M$ est donc

$$T_M = \bar{U}_M + (1 + M^{-1}) B_M.$$

Lorsque θ est un scalaire, la distribution de référence pour les intervalles d'estimation et les tests de signification est une distribution de t .

$$(\theta - \bar{\theta}_M) T_M^{-1/2} \sim t_v,$$

où la formule des degrés de liberté,

$$v = (M - 1) \{ 1 + [(1 + M^{-1}) B_M / \bar{U}_M]^{-1} \}^2$$

est fondée sur une approximation de Satterthwaite (Rubin et Schenker, 1986, et Rubin, 1986a). Le rapport de la variance intra-imputation à la variance inter-imputation \bar{U}_M / B_M donne une estimation du paramètre de population $(1 - \gamma) / \gamma$, où γ est la proportion des données manquant sur θ à cause de la non-réponse. Lorsqu'il s'agit d'un modèle de non-réponse aléatoire où il n'y a aucune covariable, γ équivaut à la proportion de données manquantes.

3.4 Niveaux de signification pour un paramètre θ à plusieurs éléments

Lorsque θ est formé de k éléments, on peut déterminer les niveaux de signification en se servant des M estimations de données complètes obtenues par itération, $\hat{\theta}_i$, et des

3. METHODE GENERALE

L'exemple de la section précédente a permis d'illustrer la méthode d'imputation multiple et la manière d'analyser les résultats de l'application de cette méthode dans un cas particulier. Nous allons maintenant décrire la méthode dans son application générale.

3.1 Méthodes d'imputation propres

Idealement, l'imputation multiple devrait être faite selon la méthode générale suivante. Pour chaque modèle considéré, les M imputations des valeurs manquantes, X^{mis} , sont autant d'itérations faites à partir de la distribution prévisionnelle a posteriori X^{mis} , chaque itération étant un prélèvement indépendant des paramètres et des valeurs manquantes selon un modèle basé sur la méthode de Bayes appliqué au mécanisme de réponse. En pratique, on peut souvent remplacer les modèles explicites par des modèles implicites comme cela a été le cas à la section 2. Les deux genres de modèles sont illustrés dans Herzog et Rubin (1983), où les auteurs font de l'imputation multiple à l'aide d'un modèle de régression explicite et d'un modèle d'appariement implicite, lequel est une modification de la méthode du hot-deck du Census Bureau.

Les méthodes qui tiennent compte de la variabilité des itérations à l'intérieur d'un modèle sont appelées *propres*, une définition précise de ces méthodes est fournie dans Rubin (1986a). L'objet fondamental des méthodes d'imputation propres est de refléter fidèlement la variabilité d'échantillonnage lorsqu'on procède à des imputations multiples dans un modèle. Prenons, par exemple, un modèle de non-réponse aléatoire, c'est-à-dire un modèle où l'on suppose que les valeurs de X des répondants et des non-répondants qui ont la même valeur de X diffèrent uniquement par hasard l'une de l'autre. Le seul fait d'imputer aléatoirement des valeurs de X à l'aide des données fournies par les répondants suffit à sous-estimer la variabilité d'échantillonnage. Cette variabilité tient à la différence aléatoire qui existe entre les valeurs de X de l'échantillon de répondants à X et les valeurs de X de la population à X . Pour avoir une idée juste de cette variabilité, il faut faire des inférences d'imputation itérative valables en vertu du mécanisme de réponse prédéterminé.

En se servant d'échantillons aléatoires simples dans un modèle de non-réponse aléatoire, Rubin et Schenker (1986) étudient l'imputation par hot-deck (qui consiste à imputer des valeurs par simple prélèvement aléatoire des données fournies par les répondants), laquelle n'est pas une méthode propre, ainsi que diverses méthodes d'imputation propres fondées sur des modèles explicites et implicites, y compris un modèle entièrement normal, la méthode bootstrap bayésienne (Rubin, 1981), et une méthode approximative bootstrap bayésienne. Cette dernière méthode (MABB) peut servir à illustrer la manière dont une méthode d'imputation intuitive comme le hot-deck aléatoire simple peut être transformée en une méthode propre.

3.2 Exemple d'application d'une méthode d'imputation propre dans un modèle de non-réponse aléatoire - MABB

Considérons un échantillon aléatoire simple de taille n avec n_R répondants et $n_{NR} = n - n_R$ non-répondants. La MABB produit M imputations itératives sans biais de la façon suivante. Pour $l = 1, \dots, M$, déterminons n valeurs possibles de X en prélevant tout d'abord parmi les n_R valeurs observées de X un échantillon aléatoire avec remise de n valeurs et en prélevant ensuite parmi ces n valeurs un échantillon aléatoire avec remise des n_{NR} valeurs manquantes de X . Comme le montrent Rubin et Schenker (1986), le fait de prélever les n_{NR} valeurs manquantes dans un échantillon de n valeurs possibles plutôt que dans l'échantillon des n_R valeurs observées produit une variance inter-imputation acceptable, du moins dans les grands échantillons. Le MABB devient une approximation de la méthode bootstrap bayésienne lorsqu'une distribution multinomiale à l'échelle sert d'approximation à une distribution de Dirichlet.

Tableau 7
Estimations par quotient et variances correspondantes pour les séries de données complètes des tableaux 3 à 6

Modèle 1		Modèle 2	
Itération		Itération	
1	2	1	2
Estimation	13.38	13.57	13.85
Variance	2.96	3.19	3.38
			3.84

Tableau 8
Valeurs combinées des estimations et des variances relatives aux séries de données complètes établies à l'aide des données des tableaux 1 et 2

Modèle 1		Modèle 2	
Estimation		Estimation	
13.48		13.98	
Variance		3.66	

où la sommation s'opère sur les unités de l'échantillon. Le tableau 7 donne les estimations et les variances se rapportant aux quatre séries de données reproduites dans les tableaux 3 à 6.

Les deux valeurs calculées pour chaque modèle peuvent être combinées afin de produire une seule inférence pour \bar{Y} par modèle. Les résultats obtenus figurent au tableau 8; l'estimation correspond à la moyenne des estimations du tableau 7 tandis que la variance liée à cette estimation a deux composantes: (i) la variance intra-imputation moyenne de l'estimation et (ii) la variance inter-imputation de l'estimation. Ainsi, pour le modèle 1, l'estimation est $(13.38 + 13.57)/2 = 13.48$; la variance intra-imputation moyenne estimée est $(2.96 + 3.19)/2$ et la variance inter-imputation estimée correspondante est $[(13.38 - 13.48)^2 + (13.57 - 13.48)^2]$. Ces deux variances sont combinées selon la formule suivante: (variance totale estimée) = (variance intra-imputation moyenne estimée) + $(1 + M^{-1}) \times$ (variance inter-imputation estimée), où le coefficient $(1 + M^{-1})$, qui multiplie l'estimation non biaisée courante de la variance inter-imputation, sert de facteur de correction pour tenir compte de l'utilisation d'un nombre fini d'imputations. L'intervalle d'estimation pour \bar{Y} , avec un seuil de confiance de 95%, est (10.0, 16.9) selon le modèle 1 et (10.2, 17.7) selon le modèle 2. En pratique, il est possible d'obtenir de meilleurs intervalles en calculant les degrés de liberté comme une fonction simple des composantes de la variance et en appliquant le seuil de confiance de 95% à une distribution de t ; si M est grand ou si la variance inter-imputation est faible par rapport à la variance totale (comme c'est le cas dans l'exemple ci-dessus), le nombre de degrés de liberté sera élevé et un seuil de confiance de 95% sera alors utilisée.

La chose essentielle à retenir de cet exemple est qu'il ne fait appel qu'à des méthodes d'analyse de données complètes. Nous n'avons qu'à soumettre chaque série de données complètes obtenue par imputation multiple à l'analyse de données complètes qui aurait été effectuée s'il n'y avait pas eu de non-réponse. Nous pouvons ensuite combiner aisément les résultats obtenus dans chaque modèle pour obtenir une inférence par modèle. Bien qu'il n'en soit pas question dans le présent document, chaque série de données complètes peut être soumise à des analyses de prévision où l'on applique des méthodes d'analyse de données complètes; plusieurs exemples de ces analyses sont contenus dans Heitjan et Rubin (1986).

Tableau 5
Série de données complètes n° 3 (modèle 2, itération 1)
établie à l'aide des données des tableaux 1 et 2

Unité		Y		X	
				moyennes	
1	8				
2	12				
3	14				
4	19				
5	16				
6	15				
7	20				
8	4				
9	18				
10	22				
15	15				
13					

Tableau 6
Série de données complètes n° 4 (modèle 2, itération 2)
établie à l'aide des données des tableaux 1 et 2

Unité		Y		X	
				moyennes	
1	10				
2	17				
3	14				
4	17				
5	16				
6	15				
7	20				
8	4				
9	18				
10	22				
15.3	15.3				
13					

2.2 Analyse des séries de données obtenues par imputation multiple

Chaque série d'imputations, c'est-à-dire chaque colonne du tableau 2, peut être intégrée aux données du tableau 1 de façon à produire une série de données complètes. Comme il y a quatre séries d'imputations, nous pouvons avoir quatre séries de données complètes; celles-ci sont reproduites dans les tableaux 3 à 6. Nous analysons chacune de ces séries de données comme s'il n'y avait jamais eu de non-réponse.

Maintenant que nous avons des données complètes, supposons que nous utilisons l'estimateur $\overline{Xy}/\overline{x}$ (estimation par le quotient) et la variance correspondante $SE^2_{\overline{Xy}/\overline{x}}$, où \overline{X} est la moyenne connue de X dans la population, par exemple 12, \overline{y} et \overline{x} sont les moyennes de Y and X dans l'échantillon aléatoire de n unités et

$$SE^2 = \sum (Y_i - X_i \overline{y}/\overline{x})^2 / [n(n - 1)]$$

Résultats de l'imputation multiple appliquée aux données du tableau 1			
Modèle 2		Modèle 3	
Itération		Itération	
1	2	1	2
Unité 2	10	14	12
Unité 4	16	14	19

Tableau 2

Série de données complètes n° 1 (modèle 1, itération 1) établie à l'aide des données des tableaux 1 et 2			
Unité		Y	
X		10	
8	10	10	10
9	10	14	14
11	3	16	16
13	4	16	16
16	5	16	16
18	6	15	15
20	9	18	18
25	10	22	22
13	moyennes	14.5	14.5

Tableau 3

Série de données complètes n° 2 (modèle 1, itération 2) établie à l'aide des données des tableaux 1 et 2			
Unité		Y	
X		10	
8	1	10	10
9	2	14	14
11	3	14	14
13	4	14	14
16	5	16	16
18	6	15	15
20	7	20	20
25	9	18	18
13	10	22	22
13	moyennes	14.7	14.7

Tableau 4

Le tableau 1 donne les valeurs observées de (Y, X) pour les 10 unités de l'échantillon; les points d'interrogation indiquent des données manquantes à cause de la non-réponse.

2.1 Imputation multiple appliquée aux valeurs manquantes

Supposons que les valeurs manquantes du tableau 1 fassent l'objet d'une double imputation suivant deux modèles distincts (c'est-à-dire deux itérations par modèle). En règle générale, on peut utiliser autant de modèles et autant d'itérations que nécessaire. Le modèle numéro 1 est un modèle de non-réponse aléatoire; ce genre de modèle est défini en termes précis dans Rubin (1976). Essentiellement, il signifie qu'un non-répondant ayant une valeur de X donnée diffère uniquement par hasard d'un répondant affichant la même valeur de X. Le modèle numéro 2 est un modèle non-aléatoire et il suppose une différence systématique entre les répondants et les non-répondants pour lesquels la valeur de X est la même. Les itérations effectuées suivant chaque modèle reposent sur une méthode simple qui se rapproche sensiblement de la méthode du hot-deck et qui, malgré certaines faiblesses, est utile pour illustrer des notions fondamentales.

Il s'agit de déterminer pour chaque unité non-répondante les deux unités répondantes qui ont le plus d'affinité avec celle-ci; on se fonde à cette fin sur les valeurs de X des unités en question. En ce qui concerne le premier non-répondant, soit l'unité 2, les deux répondants qui se rapprochent le plus de celle-ci par leur valeur de X sont les unités 1 et 3 tandis que les deux répondants qui se rapprochent le plus de la deuxième unité non-répondante, soit l'unité 4, sont les unités 3 et 5. On obtient les valeurs imputées en prélevant au hasard les valeurs de Y des deux unités répondantes déterminées de la manière décrite ci-dessus. En ce qui concerne le modèle aléatoire, nous imputons simplement la valeur de Y observée pour les deux répondants les plus près de l'unité non-répondante; les résultats figurent dans les deux premières colonnes du tableau 2. En ce qui concerne le modèle non-aléatoire, nous supposons que le biais attribuable à la non-réponse est tel que la valeur de Y pour un non-répondant aura tendance à être de 20% supérieure à la valeur de Y observée pour les répondants les plus près; les résultats obtenus figurent dans les deux dernières colonnes du tableau 2. Les valeurs de Y ont été arrondies au nombre entier le plus près. Les itérations effectuées à l'intérieur de chaque modèle permettent à l'utilisateur de faire une inférence valable suivant chaque modèle. L'utilisation de deux modèles, un modèle aléatoire et un modèle non-aléatoire, permet d'évaluer la sensibilité de l'inférence aux hypothèses concernant la non-réponse. Il est habituellement impossible de tester ces hypothèses avec les données dont nous disposons.

Tableau 1

Observations		
Unité	Y	X
1	10	8
2	?	9
3	14	11
4	?	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25

méthodes d'interview et les motifs de non-réponse qui, de ce fait, n'ont pas leur place dans des fichiers à grande diffusion, ou peuvent être des données factuelles, comme des adresses de résidence, qui ne peuvent figurer dans des fichiers à grande diffusion pour des raisons de confidentialité. Bien qu'ils soient inaccessibles aux utilisateurs de fichiers à grande diffusion, ce genre de renseignements peuvent souvent contribuer à réduire la gamme des valeurs imputées.

De même que l'imputation simple comporte des avantages indéniables, elle présente des inconvénients tout aussi indéniables en ce sens que la valeur imputée exclut toute incertitude sur la valeur à imputer. Si une valeur était vraiment appropriée, elle ne serait pas manquée. Ainsi, lorsqu'on assimile les valeurs imputées à des valeurs observées, on sous-estime systématiquement l'élément d'incertitude même en supposant que l'on connaisse les motifs exacts de la non-réponse. Fait tout aussi grave, l'imputation simple ne tient pas compte de l'élément d'incertitude qui s'ajoute lorsqu'on ne connaît pas les motifs de la non-réponse.

1.3 L'imputation multiple comme solution de rechange

L'imputation multiple, dont il a été question pour la première fois dans Rubin (1977, 1978), conserve les deux principaux avantages de l'imputation simple et corrige ses principaux inconvénients. Comme son nom l'indique, l'imputation multiple consiste à remplacer chaque valeur manquante par un vecteur de M valeurs possibles ($M \geq 2$). Les M valeurs sont ordonnées en ce sens que les premiers éléments des vecteurs servent à créer une première série de données complètes, les seconds éléments servent à créer une deuxième série de données complètes et ainsi de suite. L'imputation multiple conserve le premier grand avantage de l'imputation simple puisqu'elle prévoit l'utilisation de méthodes d'analyse de données complètes pour chaque série de données. En outre, le deuxième grand avantage de l'imputation simple, à savoir la possibilité qu'ont les recenseurs d'utiliser leur banque de données propre pour imputer des valeurs, est non seulement conservé mais accru. Grâce à l'imputation multiple, les recenseurs peuvent non seulement utiliser leur banque de données propre pour imputer des valeurs par estimation ponctuelle, mais aussi exprimer leur incertitude relativement aux valeurs à imputer. Cette incertitude prend deux formes: la variabilité d'échantillonnage, dans l'hypothèse où les motifs de non-réponse sont connus, et la variabilité, dans l'hypothèse où les motifs de non-réponse sont incertains. Pour chaque modèle de non-réponse, on effectue au moins deux imputations pour tenir compte des variabilités d'échantillonnage respectives; le fait que des imputations soient effectuées pour plus d'un modèle signifie que les motifs de non-réponse sont incertains. Les imputations successives effectuées à l'intérieur d'un même modèle sont appelées des itérations et leurs résultats peuvent être combinés de façon à permettre une inférence valable en vertu de ce modèle; il est possible de comparer les inférences obtenues suivant différents modèles afin d'évaluer la sensibilité des résultats aux motifs hypothétiques de non-réponse.

Avant d'analyser des résultats généraux à la section 3, nous allons illustrer dans la section suivante les notions fondamentales de l'imputation multiple à l'aide d'un exemple purement fictif tiré de Rubin (1986a) qui traite en profondeur l'imputation multiple. Parmi les autres ouvrages du même genre, signalons Rubin (1979, 1980, 1986b), Herzog et Rubin (1983), Li (1985), Schenker (1985), Rubin et Schenker (1986) et Heitjan et Rubin (1986).

2. EXEMPLE FICTIF DE L'IMPUTATION MULTIPLE

Supposons que nous avons prélevé un échantillon aléatoire simple de $n = 10$ unités dans une grande population. Nous cherchons à estimer la valeur Y , c'est-à-dire la moyenne de Y dans la population. Nous connaissons la valeur moyenne d'une covariable X dans la population et nous tentons par ce sondage d'enregistrer les valeurs de X et de Y pour chacune des n unités de l'échantillon.

Initiation à l'imputation multiple pour les cas de non-réponse

DONALD B. RUBIN¹

RÉSUMÉ

L'imputation multiple est une méthode de traitement de la non-réponse à une enquête qui consiste à remplacer chaque donnée manquante par un vecteur de valeurs possibles, qui reflète un certain degré d'incertitude à propos des valeurs à imputer. La description de la méthode est précédée d'un exemple simple et d'un aperçu des fondements théoriques.

MOTS CLÉS: Non-réponse à une enquête; méthodes d'imputation propres; imputation multiple.

1. INTRODUCTION

N'importe quel statisticien moins expérimenté versé dans les enquêtes par sondage sait que ce genre d'enquête n'est pas à l'abri de la non-réponse en ce sens qu'il y a toujours un certain nombre de personnes, dans les enquêtes pratiques, qui ne répondent pas à toutes les questions du questionnaire. De façon générale, les questions qui restent le plus souvent sans réponse sont celles jugées plus délicates, par exemple les questions ayant trait au revenu personnel. Comme la non-réponse se traduit par des valeurs manquantes, il n'est plus possible de calculer les statistiques que l'on cherchait à connaître à l'origine. Les personnes chargées de la collecte et de l'analyse des données cherchent donc ardemment à résoudre le problème des valeurs manquantes, et, par voie de conséquence, à retrouver l'utilisation de méthodes d'induction statistique standard.

1.1 Imputation

Il n'est donc pas surprenant de constater que l'imputation est le moyen qui a été le plus souvent utilisé pour résoudre le problème des valeurs manquantes découlant de la non-réponse. Ce moyen consiste à faire correspondre une valeur réelle à chaque valeur manquante. De nombreuses méthodes d'imputation ont été proposées, notamment celle qui consiste à appliquer à la variable manquante la moyenne des valeurs observées chez les autres répondants ou une valeur établie à partir d'un modèle fondé sur des données de l'enquête. Par exemple, lorsque la valeur manquante concerne le revenu personnel, il convient d'utiliser un modèle de régression linéaire permettant d'estimer le revenu (échelle logarithmique) à l'aide de caractéristiques démographiques telles que l'âge, le sexe, le niveau d'instruction et la profession.

1.2 Avantages et inconvénients de l'imputation simple

Un avantage évident de l'imputation est de permettre l'application de méthodes d'analyse de données complètes; ce n'est pas le seul avantage majeur qu'offre l'imputation. En effet, lorsqu'un recenseur (par exemple, le Census Bureau) effectue une imputation, il peut utiliser à cette fin des données qui ne sont pas connues des analystes de l'extérieur, comme les spécialistes des sciences sociales dans les universités, qui ont accès à des fichiers à grande diffusion. Les renseignements qui sont à l'usage exclusif du recenseur peuvent concerner les

¹ Donald B. Rubin, Université Harvard, Département de statistique, Science Center, 1 Oxford Street, Cambridge, Massachusetts, 02138.

Si les valeurs de $(f(x_i))^{1/2} \alpha_i$ sont constantes pour $i = 1, \dots, N$, il est clair d'après l'équation ci-dessus qu'il faudra prélever un échantillon relativement plus grand dans la strate et que la probabilité de réponse devra être moins élevée si l'on veut obtenir une répartition optimale. De fait, nous aurions dans un tel cas

$$E(n_j) = E(n')/k,$$

où n_j est la taille de l'échantillon effectif s_j prélevé dans la strate P_j , $j = 1, \dots, k$.

BIBLIOGRAPHIE

BREWER, K.R.W. (1963). Ratio estimation in finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

CASSEL, C.M., SARANDA, C.E., and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problems. Dans *Incomplete Data in Sample Surveys* 3, (colligé par W.G. Madow et Ingram Olkin), Academic Press: New York, 143-160.

GODAMBE, V.P. (1982). Estimation in survey sampling: Robustness and optimality. *Journal of the American Statistical Association*, 77, 393-403.

GODAMBE, V.P. (1986). Quasi-score function, quasi-observed Fisher information and conditioning in survey sampling (non publié).

GODAMBE, V.P. and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Institute Review* (à paraître).

HAJEK, J. (1971). Contribution to discussion of paper by D. Basu. Dans *Foundations of Statistical Inference* (colligé par V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston, 236.

HARTLEY, H.O. (1946). Discussion of paper by F. Yates. *Journal of the Royal Statistical Society* Série A, 109, 37.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-589.

Selon (A), nous maintenons constante la taille moyenne de l'échantillon s car $E|s| = \sum_N^i \pi_i$. Selon (B), nous maintenons constante la taille moyenne de l'échantillon effectif s' car $E|s'| = \sum_N^i \pi_i q_i$. Or, comme les valeurs de q_i sont fixes, $E(h^{**})^2$ peut être minimisée selon les conditions respectives suivantes:

$$(A): \pi_i \propto \left\{ \frac{q_i}{f(x_i)^{1/2} \alpha_i} \right\}$$

$$(B): \pi_i \propto \frac{q_i}{(f(x_i)^{1/2} \alpha_i)}$$

En désignant par n' la taille de l'échantillon effectif s' , c'est-à-dire $|s'| = n'$, nous tirons de (B) en (3.4) l'équation

$$\pi_i = \frac{(f(x_i)^{1/2} \alpha_i)}{\{\sum_N^i (f(x_i)^{1/2} \alpha_i)\}} \frac{q_i}{E(n')}, \quad i = 1, \dots, N. \tag{3.5}$$

En outre, pour un plan de sondage à taille d'échantillon fixe tel que

$$\text{Probabilité}\{s: |s| \neq n\} = 0,$$

nous déduisons de (3.5).

$$\sum_n \pi_i = n = \sum_N^i \frac{(f(x_i)^{1/2} \alpha_i)}{\{\sum_N^i (f(x_i)^{1/2} \alpha_i)\}} \frac{1}{q_i} E(n'). \tag{3.6}$$

Lorsque dans l'équation ci-dessus toutes les probabilités de réponses q_i , $i = 1, \dots, N$ sont égales, par exemple $q_i = q$, $i = 1, \dots, N$,

$$n = E(n')/q; \tag{3.7}$$

si, par exemple, $q = 1/2$, la taille de l'échantillon (initial) s devra être le double de l'espérance mathématique de la taille de l'échantillon effectif (s') !

Supposons maintenant que la population d'enquête P est divisée en strates P_j , $j = 1, \dots, k$ de sorte que les probabilités de réponse dans chaque strate soient constantes, c'est-à-dire que l'équation (2.7) soit satisfaite. Pour un plan de sondage stratifié qui consiste à prélever un échantillon de taille n_j dans la strate P_j , $j = 1, \dots, k$ nous tirons de (3.5)

$$n_j = \frac{E(n')}{q_j} \frac{\sum_{i \in P_j} (f(x_i)^{1/2} \alpha_i)}{\sum_{i \in P} (f(x_i)^{1/2} \alpha_i)}, \quad j = 1, \dots, k. \tag{3.8}$$

ne tenant pas compte des séries de valeur nulle. De plus, étant donné le nombre de réponses dans n_j de la strate P_j , la probabilité de $i \in s_j^*$ est $(n_j/N_j)(n_j/n_j) = (n_j/N_j)$. Par conséquent, pour toute fonction d'estimation $h_1 \in H_1$ dans (2.8), nous avons par la thèse 2.3 l'inéquation.

$$E\{(h_1^*)^2 | n_j, j = 1, \dots, k\} \leq E\{(h_1)^2 | n_j, j = 1, \dots, k\}, \quad (2.11)$$

h_1^* étant définie en (2.9). On démontre le théorème 2.4 en calculant les espérances mathématiques de part et d'autre de l'inéquation (2.11) pour les variations de $n_j, j = 1, \dots, k$.

On peut interpréter intuitivement la fonction d'estimation optimale h_1^* définie en (2.9) de la façon suivante: si les probabilités de réponse $q^{(j)}, j = 1, \dots, k$ en (2.7) étaient connues, alors, selon le théorème 2.3, la fonction d'estimation optimale pour le plan de sondage p_0 serait donnée par

$$h'' = \sum_k \sum_{i \in s_j^*} (y_i - \theta x_i) \alpha_i / \left(\frac{N_j}{n_j} q^{(j)} \right).$$

Or, lorsque $q^{(j)}$ est inconnue (ce qui est effectivement le cas dans le théorème 2.4), nous l'estimons par la formule $(n_j/n_j), j = 1, \dots, k$. Si nous substituons cette expression à $q^{(j)}$ dans h'' , nous obtenons la fonction d'estimation h_1^* définie en (2.9).

Les estimations découlant de la résolution des équations $h_1'' = 0, h_1''^* = 0$ et $h_1^* = 0$ définies respectivement en (2.5), (2.6) et (2.9) ont déjà été proposées par plusieurs auteurs selon des critères de vraisemblance. À ce sujet, l'ouvrage de Cassel et coll. (1983) mérite d'être cité. L'hypothèse relative aux probabilités de réponse (2.4) semble avoir évolué peu à peu dans les ouvrages de statistique. Hartley (1946) est une source intéressante à cet égard.

3. PROBABILITÉS D'INCLUSION OPTIMALE

À ce stade-ci, il convient de souligner que le caractère optimal de la fonction d'estimation h''^* définie en (2.6) a été établi suivant le modèle de superpopulation (1.2), lequel ne définit pas la fonction de variance. Il faudrait pourtant définir cette fonction dans le modèle précité afin de connaître les probabilités d'inclusion optimale. Nous posons par hypothèse

$$E(y_i - \theta x_i)^2 = \sigma^2 f(x_i), \quad i = 1, \dots, N, \quad (3.1)$$

où f est une fonction connue de x , et σ^2 peut être inconnue. Si maintenant nous appliquons l'équation (3.1) à la fonction d'estimation h''^* définie en (2.6), nous obtenons

$$E(h''^*)^2 = \sum_N \frac{E(y_i - \theta x_i)^2 \alpha_i^2}{\epsilon(y_i - \theta x_i)^2 \alpha_i^2} = \sigma^2 \sum_N \frac{\pi_i q_i}{f(x_i) \alpha_i^2}. \quad (3.2)$$

Dans l'équation ci-dessus, les probabilités de réponse q_i définies en (2.4) sont des valeurs données (fixes). Par ailleurs, on peut établir les probabilités d'inclusion optimale d'un plan de sondage en minimisant $E(h''^*)^2$ dans (3.2) moyennant la restriction (A) ou (B).

$$(A): \sum_N \pi_i = \text{constante},$$

$$(B): \sum_N \pi_i q_i = \text{constante}. \quad (3.3)$$

Lorsque $s \equiv s'$, c'est-à-dire lorsqu'il n'y a aucun non-répondant, doit-on encore estimer h^* par $h'^* = h''^*$? Il est évident que non et Godambe (1986) le montre par des conditions appropriées. On tend à se poser la même question dans les cas où il y a très peu de non-répondants et là encore, on peut trouver une réponse à cette question en posant des conditions appropriées. En définitive, le caractère optimal formel de la relation $h'^* = h''$ donne à croire que cette relation est utile et qu'elle est de nature à produire une bonne estimation lorsque le taux de non-réponse est élevé et que les valeurs relatives de q_i sont connues. Toutefois, elle est loin d'être aussi efficace lorsque le taux de non-réponse est peu élevé; dans un tel cas, cette relation gagnerait à être améliorée car on pourrait alors en tirer naturellement des conditions. Il est encore plus important de poser les hypothèses appropriées lorsqu'on suppose des probabilités de réponse $q^{(U)}$, $j = 1, \dots, k$ inconnues. Désignons par p_0 le plan de sondage stratifié qui consiste à prélever dans le strate \mathbf{P}_j un échantillon aléatoire simple (sans remise) de taille n_j , $j = 1, \dots, k$. Comme en (2.3), nous définissons maintenant la classe des fonctions d'estimation sans biais $h_1(X_{s,s'})$

Supposons que la population d'enquête \mathbf{P} est divisé en k strates \mathbf{P}_j de taille N_j , $j = 1, \dots, k$. Supposons de plus que les probabilités de réponse dans chaque strate sont constantes, c'est-à-dire

$$q_i = q^{(U)} \text{ for all } i \in \mathbf{P}_j; j = 1, \dots, k. \tag{2.7}$$

Contrairement à (2.4), où l'on supposait que les probabilités de réponse étaient connues, l'expression (2.7) suppose des probabilités de réponse $q^{(U)}$, $j = 1, \dots, k$ inconnues. Désignons par p_0 le plan de sondage stratifié qui consiste à prélever dans le strate \mathbf{P}_j un échantillon aléatoire simple (sans remise) de taille n_j , $j = 1, \dots, k$. Comme en (2.3), nous définissons maintenant la classe des fonctions d'estimation sans biais $h_1(X_{s,s'})$

$$H_1(p_0) = \{h_i: E(h_i - \bar{g}) = 0 \text{ pour toutes les valeurs de } y, \theta \text{ et } q^{(U)}, j = 1, \dots, k\}, \tag{2.8}$$

où $q^{(U)}$ est défini comme en (2.7). Posons $s_j' = s' \cap \mathbf{P}_j$ et $|s_j'| = n_j$, c'est-à-dire la taille de l'échantillon de répondants de la strate \mathbf{P}_j , $j = 1, \dots, k$.

Théorème 2.4. Pour le plan de sondage p_0 dans la classe $H_1(p_0)$ des fonctions d'estimation définie en (2.8), selon le modèle de superpopulation (1.2), $E(h_1^2)$ est minimisée pour $h_1 = h_1^*$ où

$$h_1^* = \sum_k \sum_{i \in s_j'} (y_i - \theta x_i) \alpha_i / (N_j/n_j); \tag{2.9}$$

et d'autres termes, h_1^* est la fonction d'estimation optimale dans $H_1(p_0)$.

Preuve. La distribution d'échantillonnage des données $X_{s,s'}$ définies en (2.1) dépend non seulement du vecteur de population inconnu y , mais aussi du paramètre inconnu $q^{(U)}$, $j = 1, \dots, k$. Or, pour chaque vecteur y donné, le paramètre statistique n_j , $j = 1, \dots, k$ est *entièrement suffisant* pour le paramètre $q^{(U)}$, $j = 1, \dots, k$. Donc, pour des valeurs données de y et de θ dans (2.8),

$$[E(h_1 - \bar{g}) = 0, \text{ pour tout } q^{(U)}, j = 1, \dots, k] \\ \Rightarrow E\{(h_1 - \bar{g})|n_j, j = 1, \dots, k\} = 0, \tag{2.10}$$

MÉCANISME DE RÉPONSE: Si l'individu i de la population d'enquête \mathbf{P} est inclus dans l'échantillon s

la probabilité que i fournisse une réponse est q_i
 et la probabilité qu'il ne fournisse pas de réponse est $1 - q_i$

$i = 1, \dots, N$; nous posons que $q_i > 0, i = 1, \dots, N$.

À l'aide du mécanisme de réponse $\mathbf{q} = (q_1, \dots, q_N)$ défini en (2.4), nous pouvons définir entièrement la classe $H'(p, \mathbf{q}, s)$ dans (2.2) comme $H'(p, \mathbf{q}, s)$ et la classe $H''(p, \mathbf{q})$ dans (2.3) comme $H''(p, \mathbf{q})$.

Le théorème 2.1 ci-dessous est l'application du cas (I) tandis que les théorèmes 2.2, 2.3 et 2.4 sont l'application du cas (II).

Théorème 2.1. Pour tout plan de sondage p satisfaisant à l'inéquation (1.11) et pour tout échantillon s dans la classe $H'(p, \mathbf{q}, s)$ des fonctions d'estimation (équation 2.2), selon le modèle de superpopulation (1.2), $eE\{(h')^2 | s\}$ est minimisée pour $h' = h'^*$ où

$$h'^* = \sum_{i \in s'} (y_i - \theta x_i) \alpha_i / \pi_i q_i; \quad (2.5)$$

en d'autres termes, h'^* est la fonction d'estimation optimale dans $H'(p, \mathbf{q}, s)$.

Preuve. Comme il a été souligné dans la section 1, le caractère optimal de h^* dans (1.12) tient pour les plans de sondage où la taille de l'échantillon est aléatoire et pour n importe quelle valeur de $\alpha_p, i = 1, \dots, N$ dans (1.3). On peut donc faire la preuve du théorème 2.1 en remplaçant dans la théorème 1.1 la population \mathbf{P} par s et α_i par $\alpha_i / \pi_i, i \in s$ et en précisant que les probabilités d'inclusion sont désormais désignées $q_i, i \in s$.

Théorème 2.2. Soit H'' la sous-classe de H'' dans (2.3) de telle sorte que toute fonction d'estimation $h''(x_{s,s'})$ dans H'' ne dépend que de s dans (s, s') . Alors, pour tout plan de sondage p satisfaisant à l'inéquation (1.11), dans la classe $H''(p, \mathbf{q})$, selon le modèle de superpopulation (1.2), $eE\{(h'')^2\}$ est minimisée pour $h'' = h''^*$ où

$$h''^* = \sum_{i \in s'} (y_i - \theta x_i) \alpha_i / \pi_i; \quad (2.6)$$

en d'autres termes, h''^* est la fonction d'estimation optimale dans $H''(p, \mathbf{q})$.

Preuve. Elle découle directement du théorème 1.1, dans lequel il suffit de remplacer s par s' et les probabilités d'inclusion π_i par $\pi_i q_i, i = 1, \dots, N$.

Théorème 2.3. La fonction d'estimation h''^* définie en (2.6) est optimale dans toute la classe $H''(p, \mathbf{q})$ définie en (2.3). En d'autres termes, l'application du théorème 2.2 ne se limite pas à la sous-classe H'' de H'' .

Preuve. Pour n importe quelle probabilité de réponse \mathbf{q} en (2.4) et n importe quel plan de sondage p , le paramètre statistique $\{(i, y_i) : i \in s'\}$ est suffisant pour le vecteur de population y . En termes plus précis, si nous nous reportons aux équations (1.1) et (2.1), la probabilité conditionnelle $\text{Prob}(x_{s,s'} | x_{s'}, y)$ est indépendante de y . Ainsi, pour toute fonction d'estimation $h'' \in H''(p, \mathbf{q})$ dans (2.3), nous avons la fonction d'estimation $E(h'' | x_{s'}) = h'' \in H''$ et $eE(h'')^2 < eE(h'')^2$, ce qui démontre le théorème 2.3.

Godambe et Thompson: Résultats optimaux

optimaux, peu importe que la taille de l'échantillon soit fixe ou variable. Autrement dit, on peut obtenir des résultats optimaux dans des plans de sondage où la taille de l'échantillon est aléatoire; ce genre de plans de sondage se présente souvent dans les cas de non-réponse analysés plus loin.

2. NON-RÉPONSE ET RÉSULTATS OPTIMAUX

Supposons qu'un échantillon s est prélevé dans la population d'enquête P au moyen d'un plan de sondage p . Supposons aussi qu'à cause de la non-réponse, nous connaissons uniquement les valeurs de la variable y_i relatives aux sous-ensembles $s' \subset s$ — s' étant les non-répondants. Ainsi, au lieu d'être représentés par X_s (équation 1.1), les données correspondantes sont désormais représentées par l'expression

(2.1) $X_{s,s'} = (s, s' \{ (i, y_i) : i \in s' \})$.

Nous pouvons alors considérer deux problèmes d'estimation:

(I) En l'absence de non-réponse, c'est-à-dire si nous connaissons toutes les données X_s définies en (1.1), nous estimerions le paramètre de population d'enquête θ_N défini en (1.4) en résolvant l'équation d'estimation optimale définie en (1.12), c'est-à-dire $h^* = 0$. Lorsque les données hypothétiques sont remplacées par $X_{s,s'}$ défini en (2.1), on peut chercher à estimer h^* au moyen d'une fonction h' ($X_{s,s'}$). Cela est en accord avec une proposition de Rubin (1976). Conformément à l'équation (1.7), nous définissons la classe des fonctions d'estimation sans biais h' (pour h^* , étant donné l'échantillon s) comme

(2.2) $H'(p, s) = \{h' : E(h' - h^* | s) = 0, \text{ pour toutes les valeurs de } y \text{ et } \theta\};$

le point (.) dans H' indique que la classe H' ne peut être définie qu'une fois que le mécanisme de réponse a été précisé. De nouveau, nous définissons h'^* comme la fonction d'estimation optimale dans H' (équation 2.2) si $h'^* \in H'$ et si, aux termes du modèle (1.2), $E(h'^*)^2 \leq E(h^*)^2$ pour tout $h' \in H'$.

(II) Par ailleurs, nous pourrions chercher à estimer *directement* (c'est-à-dire, sans estimer h^* comme en (I)) le paramètre de population d'enquête θ_N à l'aide des données $X_{s,s'}$. Conformément à l'équation (1.7), nous définissons la classe des fonctions d'estimation sans biais h'' ($X_{s,s'}$) comme

(2.3) $H''(p, .) = \{h'' : E(h'' - g) = 0, \text{ pour toutes les valeurs de } y \text{ et } \theta\};$

comme dans le cas précédent, le point (.) dans H'' indique que la classe H'' ne peut être définie qu'une fois que le mécanisme de réponse a lui-même été défini. À nouveau, nous définissons h'' comme la fonction d'estimation optimale dans H'' si $h''^* \in H''$ et si, selon le modèle (1.2), $E(h''^*)^2 \leq E(h^*)^2$ pour toutes les fonctions d'estimation $h'' \in H''$. Dans les expressions $H'(p, s)$ et $H''(p, .)$, le mécanisme de réponse (.) n'est pas défini. Nous nous attachons maintenant à le définir.

Théorème 1.1. (Godambe et Thompson 1986). Pour tout plan de sondage p qui satisfait à l'inéquation (1.11), aux termes du modèle (1.2), dans la classe $H(p)$ de toutes les fonctions d'estimation sans biais (équation (1.7)), la fonction optimale h^* , c'est-à-dire h^* satisfaisant à l'inéquation (1.8), est définie par

$$(1.12) \qquad h^* = \sum_{i \in s} (y_i - \theta x_i) \alpha_i / \pi_i$$

π_i étant la probabilité d'inclusion définie en (1.10). L'équation d'estimation optimale s'exprime donc:

$$(1.13) \qquad \sum_{i \in s} (y_i - \theta x_i) \alpha_i / \pi_i = 0.$$

L'estimateur $\hat{\theta}_s$ du paramètre de population d'enquête θ_N défini en (1.4) et du paramètre de superpopulation θ défini en (1.2) est défini comme

$$(1.14) \qquad \hat{\theta}_s = \frac{\sum_{i \in s} y_i \alpha_i / \pi_i}{\sum_{i \in s} x_i \alpha_i / \pi_i}.$$

Brewer (1963) et Hajek (1971) ont déjà proposé cet estimateur selon des critères de "vraisemblance".

Pour faire le lien entre le théorème 1.1 et les résultats optimaux obtenus antérieurement (voir, par exemple, Godambe, 1982) nous posons $\alpha_i \equiv 1$ dans l'équation (1.3) et, partant, dans l'équation (1.2). De plus, nous considérons un modèle de superpopulation que l'on tire de (1.2) en posant $\theta = \theta_0$, celle-ci étant une valeur définie. Alors, pour tout plan de sondage avec des probabilités d'inclusion π_i satisfaisant à l'inéquation (1.11), dans la classe de tous les estimateurs sans biais de θ_N du plan de sondage (dans (1.4)), $\alpha_i = 1, i = 1, \dots, N$), l'espérance mathématique de la variance du plan de sondage dans ce modèle de superpopulation est minimisée pour l'estimation

$$(1.15) \qquad e = \frac{1}{X} \left\{ \sum_{i \in s} \frac{y_i}{x_i - \theta_0 x_i} \pi_i + \theta_0 \sum_{i=1}^N x_i \right\}$$

où $X = \sum_{i=1}^N x_i$. Le caractère optimal de l'estimation e à $\theta = \theta_0$ vaut pour toutes les valeurs de θ si et seulement si le plan de sondage est tel que

$$(1.16) \qquad \text{Probability} \left\{ s: \left(\sum_{i \in s} \frac{x_i}{\pi_i} - \sum_{i=1}^N x_i \right) = 0 \right\} = 1.$$

Or, lorsque le plan de sondage satisfait à la condition (1.16), θ_s dans (1.14) est égal à e dans (1.15). Ainsi, le théorème 1.1 explique tous les résultats optimaux obtenus antérieurement, mais il ne s'arrête pas là. Dans beaucoup de cas, notamment lorsqu'il s'agit de plans de sondage où $\pi_i \propto x_i$, la condition (1.16) suppose un plan de sondage où la taille de l'échantillon est fixe. Par ailleurs, selon le théorème 1.1, on peut obtenir des résultats

C'est-à-dire,

$$\theta_N = \sum_N^i y_i \alpha_i / \sum_N^i x_i \alpha_i. \tag{1.4}$$

Le paramètre θ_N se rattache au modèle (1.2) par l'équation

$$E \tilde{g} = 0. \tag{1.5}$$

Toute fonction réelle h des données X_s dans (1.1) et du paramètre θ est désignée *une fonction d'estimation sans biais* des paramètres θ_N et θ si

$$E(h - \tilde{g}) = 0 \text{ pour toutes les valeurs de } y \text{ et } \theta, \tag{1.6}$$

E étant l'espérance mathématique s'appliquant au plan de sondage p utilisé pour former l'échantillon s . Étant donné les équations (1.5) et (1.6), nous disons que la solution de l'équation

$$h(X_s, \theta) = 0,$$

pour l'ensemble de données X_s est une estimation des paramètres θ et θ_N définis respectivement par (1.2) et (1.4). Pour la fonction \tilde{g} dans (1.4), suivant le plan de sondage p , posons $H(p)$ comme la classe de toutes les fonctions d'estimation sans biais h . C'est-à-dire

$$H(p) = \{h: E(h - \tilde{g}) = 0 \text{ pour toutes les valeurs de } y \text{ et } \theta\}. \tag{1.7}$$

Alors, nous disons qu'une *fonction d'estimation* $h^* \in H(p)$ est optimale si

$$E(E(h^*)^2) \leq E(E(h)^2) \text{ pour tout } h \in H(p) \tag{1.8}$$

(Godambe et Thompson 1986). De plus, lorsque l'inéquation (1.8) est satisfaite,

$$h^* = 0 \tag{1.9}$$

est considérée comme *l'équation d'estimation optimale* pour l'estimation du paramètre θ_N défini par (1.3) et (1.4).

Pour le plan de sondage p utilisé pour le prélèvement de l'échantillon s , posons $\pi_i, i = 1, \dots, N$ comme les probabilités d'inclusion. En d'autres termes,

$$\pi_i = \sum_{s \ni i} p(s), i = 1, \dots, N, \tag{1.10}$$

où s i désigne tous les échantillons s qui renferment l'individu i . Nous posons par hypothèse

$$\pi_i > 0, i = 1, \dots, N. \tag{1.11}$$

Résultats optimaux en situation de non-réponse

V. P. GODAMBE et M. E. THOMPSON¹

RÉSUMÉ

L'application de fonctions d'estimation optimales aux enquêtes par sondage (Godambe et Thompson 1986) produit des résultats optimaux en situation de non-réponse.

MOTS CLÉS: Fonction d'estimation optimale; non-réponse.

1. INTRODUCTION ET GÉNÉRALITÉS

Une enquête par sondage est habituellement caractérisée par une population \mathbf{P} de N individus i ; $\mathbf{P} = \{i: i = 1, \dots, N\}$. À chaque individu i correspond une valeur réelle y_i . Le vecteur $\mathbf{y} = (y_1, \dots, y_p, \dots, y_N)$ est appelé vecteur de population. Tout sous-ensemble s de \mathbf{P} est appelé échantillon. Posons $S = \{s\}$. On appelle plan de sondage une distribution de probabilité p sur S . On prélève un échantillon s au moyen d'un plan de sondage p et on détermine les valeurs y_i ; i es par une enquête. En l'occurrence, les données sont représentées par x_s où

$$(1.1) \quad x_s = \{s, (i, y_i): i \in s\}.$$

À l'aide des données définies par (1.1), on cherche à estimer un paramètre de population d'enquête θ_N , qui est une fonction réelle particulière du vecteur de population \mathbf{y} ; $\theta_N = \theta_N(\mathbf{y})$. À cet égard, nous supposons un modèle de superpopulation suivant lequel y_1, \dots, y_N sont indépendantes et, pour certaines valeurs connues x_i de covariables $i = 1, \dots, N$,

$$(1.2) \quad e(y_i - \theta x_i) = 0, \quad i = 1, \dots, N,$$

étant l'espérance mathématique relative au modèle. Dans le modèle (1.2), θ est le paramètre de régression *inconnu* habituel, l'espérance mathématique étant supposée maintenir les valeurs de x_i fixes. L'ordonnée à l'origine du modèle de régression n'est pas précisée dans (1.2) car il est souvent possible de l'éliminer par une méthode de stratification appropriée (Godambe, 1982). Il est à noter que le modèle (1.2) ne définit pas de fonction de variance.

Selon Godambe et Thompson (1986), pour des nombres particuliers α_i , $i = 1, \dots, N$, nous définissons le paramètre de population d'enquête θ_N comme la solution de l'équation

$$(1.3) \quad \bar{g} = \sum_{i=1}^N (y_i - \theta x_i) \alpha_i = 0.$$

¹ V. P. Godambe et M. E. Thompson, Département de la statistique et des sciences actuarielles, Université de Waterloo, Waterloo (Ontario), Canada, N2L 3G1

- PLATEK, RICHARD et GRAY, G.B. (1985). Méthodes de compensation de la non-réponse. *Techniques d'enquête*, 11, 1-14.
- WISEMAN, FREDERICK, et PHILIP McDONALD (1978). The nonresponse problem in consumer telephone survey. Rapport n° 78-116, Marketing Science Institute, Cambridge, (Mass.).
- WISEMAN, FREDERICK, et PHILIP McDONALD (1980). Toward the development of industry standards for response and nonresponse rates. Rapport n° 80-101, Marketing Science Institute, Cambridge, (Mass.).

Une autre question relative à l'uniformisation des définitions des taux de réponse ou de non-réponse, c'est de savoir quelles normes l'expérience permettra d'établir pour des enquêtes, des spécialités et des méthodes d'interview données. Par exemple, d'après l'équation (3.2), le taux de réponse dans l'EPA est censé se situer entre 93 et 95%, sauf durant l'été, où il est légèrement inférieur. Sur les 5 à 7 points de pourcentage associés à la non-réponse, environ 1 point de pourcentage est sensé être attribuable à des refus. Depuis que cette enquête existe, les taux globaux ont été remarquablement stables.

On a constaté (voir Platek, 1977) que les enquêtes portant sur des questions financières avaient un taux de réponse moins élevé (un taux de non-réponse plus élevé) que les enquêtes sur des questions d'un autre ordre. Le taux de non-réponse pour les enquêtes financières semble se situer autour de 25% tandis que le taux de non-réponse pour la plupart des autres enquêtes fluctue entre 10 et 15%. De même, les enquêtes par téléphone semblent avoir un taux de non-réponse légèrement supérieur (de 2 à 3%) à celui des enquêtes sur place pour un même sujet. Il est donc possible de se fonder sur l'expérience pour définir un objectif standard pour les enquêtes portant sur un sujet donné et mettant en application une méthode d'interview donnée. On a constaté dans certains ouvrages, par exemple Wiseman et McDonald (1980), que les avis étaient partagés sur la manière de définir et de mesurer la non-réponse. Il semble donc que l'on doive composer avec les diverses définitions et les divers termes proposés et tenter d'établir des relations entre eux dans diverses conditions d'enquête. Nous avons tenté en vain de résoudre ce problème d'hétérogénéité. Une étude sérieuse en ce sens ne peut être réalisée que grâce à un examen approfondi des enregistrements d'enquête, lequel examen n'est possible que si des données exactes ont été enregistrées. Il arrive souvent dans les enquêtes par téléphone ou par la poste, surtout lorsqu'il s'agit d'échantillons par la méthode des quotas, que l'on substitue de nouvelles unités aux unités non-répondantes et qu'on les traite comme s'il s'agissait des unités de l'échantillon original. La qualité de l'enquête paraît donc supérieure à ce qu'elle est en réalité à cause de la dissimulation du biais dû à la non-réponse. Il est donc nécessaire de déterminer comment nous traiterons les non-répondants et comment nous évaluerons la non-réponse, l'intégralité, etc., avant de procéder à la collecte des données, de manière que nous n'ayons pas à le faire pendant ou après l'enquête.

BIBLIOGRAPHIE

- CANNELL, CHARLES (1978). Discussion of response rates. Health Survey Research Methods Conference, Department of Health, Education and Welfare, Publication n° (PHS) 79-3207.
- HAUCK, MATTHEW (1974). Planning field operations. Dans *Handbook of Marketing Research* (Robert Ferber ed.), New York: McGraw-Hill, 147-159.
- KALTON, GRAHAM (1981). *Compensating for missing survey data*. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.
- KLECKA, W.R. and A.J. TUCHFARBER (1979). Random digit dialing: A comparison to personal surveys. *Public Opinion Quarterly* (Spring), 105-114.
- KVIZ, FREDERICK J. (1977). Toward a standard definition of response rate. *Public Opinion Quarterly* (Summer), 265-267.
- LINDSTRÖM, HAKAN (1983). Non-response errors in sample surveys. Ural, Nummer 16, Skriftserie utgiven av Statistiska Centralbyrån, Statistics Sweden, Stockholm.
- O'NEIL, MICHAEL J., GROVES, ROBERT M., et CANNELL, CHARLES F. (1979). Telephone interview introductions and refusal rates: Experiments in increasing respondent cooperation. Document présenté à la conférence de 1979 de l'American Statistical Association, Washington, D.C.
- PLATEK, RICHARD (1977). Some factors affecting non-response. Document présenté à l'Institut international de statistique, New Delhi, décembre.

le nombre d'unités répondantes (admissibles ou concernées par la question "y"/non-concernées par la question "y")/(répondent à la question "y"/refusent de répondre à la question "y"), etc. La plupart des taux se rattachant aux questionnaires, autres que les taux de contact, devraient avoir un équivalent qui se rattache aux questions; cet équivalent serait facilement obtenu en faisant les substitutions appropriées dans les formules de taux. Il peut être plus difficile de connaître les motifs de non-réponse à une question que les motifs de non-réponse au questionnaire puisqu'il arrive souvent que la non-réponse à une question ne soit constatée qu'à l'aide d'un programme de contrôle et d'imputation.

f) *Taux pondérés et taux par caractéristique*

Dans le cas d'un échantillon dont les unités sont affectées de poids de sondage différents Π_i^{-1} , comme dans l'échantillonnage avec probabilité proportionnelle à la taille (*ppt*), on pondère tous les taux définis ci-dessus en appliquant, dans les équations correspondantes, le poids de sondage Π_i^{-1} à l'indice de sélection i . Dans le cas d'échantillons auto-pondérés dans un secteur ou une catégorie pour lesquels les taux sont déterminés, les poids de sondage sont superflus. Cependant, au dernier degré de l'échantillonnage *ppt*, les grandes unités ont habituellement tendance à répondre plus rapidement que les petites, de sorte que les taux de réponse pondérés sont généralement inférieurs aux taux non-pondérés fondés sur les formules de dénombrement des unités du tableau 1, compte tenu du fait que les poids de sondage appliqués aux grandes unités sont inférieurs à ceux appliqués aux petites unités. Les taux de réponse pondérés servent à estimer la proportion de la population qui aurait répondu à l'enquête dans des conditions semblables tandis que les taux de réponse non-pondérés donnent cette estimation seulement pour l'échantillon ou le sous-échantillon relatif à une catégorie ou à un secteur particulier.

En permettant d'estimer le taux de non-réponse pour la population entière plutôt que pour l'échantillon, comme le font les taux non-pondérés, les taux pondérés risquent de produire des renseignements erronés sur la qualité des données puisqu'ils peuvent fausser la distribution des caractéristiques dans l'échantillon. En revanche, comme avec les taux pondérés les unités s'ajoutent à des niveaux de population plutôt qu'à des niveaux d'échantillon, il est possible d'estimer le taux pour un recensement qui se déroulerait dans des conditions comparables. Dans certaines circonstances, les taux de réponse pondérés peuvent servir de correction de recension de poids pour compléter l'échantillon dans les cellules de correction. Lorsqu'on définit le taux de réponse pour une caractéristique, on tient compte de la réponse y_i fournie par les répondants, de la valeur imputée z_{ij} pour la non-réponse à la question et de la valeur imputée z_j pour la non-réponse au questionnaire, laquelle valeur correspond habituellement à la moyenne des répondants dans une cellule de correction. Si une valeur auxiliaire X_j peut être associée à toutes les unités, que celles-ci répondent ou non, on peut facilement calculer un taux de réponse pour une caractéristique x et s'en servir comme facteur de correction de poids lorsqu'il y a une corrélation étroite entre x et y . Le taux de réponse pour la caractéristique y , pondéré par Π_i^{-1} ou non-pondéré, peut servir à analyser le biais probable dû à la non-réponse si on le compare aux taux de réponse pondérés ou non-pondérés qui sont fondés sur le dénombrement d'unités.

4. CONCLUSION

Il semble difficile d'uniformiser les définitions de taux à cause de la diversité des usages et des études faites sur la non-réponse et à cause du soin qu'exige des recenseurs la tenue d'archives. Dans la mesure où les taux sont clairement définis et qu'ils sont appliqués convenablement dans une analyse, il peut n'être plus si important d'avoir des définitions uniformes de taux pour tous les genres d'enquêtes et de méthodes de collecte des données. Dans chaque cas, toutefois, il conviendrait de définir soigneusement le taux utilisé et de formuler clairement le but et le motif de son utilisation.

Le taux de contact permet d'évaluer la capacité de l'organisme d'enquête ou de ses interviewers de contacter les répondants, qu'ils réussissent ou non à obtenir leur collaboration. Dans l'EPA, le taux de contact pour les logements occupés se situe autour de 96% à chaque mois.

d) *Taux de refus (taux de non-refus)*

Hauck (1974) et Wiseman et McDonald (1980) définissent respectivement le taux de refus comme suit:

$$F_1 = \frac{\text{nombre de refus}}{\text{total des interviews achevées et des refus}}$$
$$= \frac{\sum_i t_i e_i (1 - \delta_i) r_i / [\sum_i t_i e_i \delta_i + \sum_i t_i e_i (1 - \delta_i) r_i]}{(\text{case 18}) / [(\text{case 9}) + (\text{case 10}) + (\text{case 18})] = 1 - R_{(3)}.$$
(3.6)

$$F_2 = \frac{\text{nombre de refus}}{\text{total des unités choisies}}$$
$$= \frac{\sum_i t_i (1 - \delta_i) r_i / \sum_i t_i}{(\text{case 18}) / (\text{case 4})}.$$
(3.7)

Si l'on applique les critères d'admissibilité, le taux défini en (3.7) devient:

$$F_3 = \sum_i t_i e_i (1 - \delta_i) r_i / \sum_i t_i e_i$$
(3.8)
$$= (\text{case 18}) / (\text{case 8}), \text{ où } e_i \text{ est définie comme dans l'équation (3.5).}$$

Le taux de refus indique dans quelle mesure l'organisme d'enquête ou l'interviewer est incapable d'obtenir la collaboration des unités choisies pour l'enquête par rapport à l'ensemble des unités contactées (3.6), par rapport à l'ensemble de l'échantillon (3.7) ou par rapport à l'échantillon admissible (3.8). On se sert de l'équation (3.6) pour calculer un taux de refus "pur", c'est-à-dire un taux qui ne tient pas compte des unités non-contactées, ce contre quoi les interviewers ne peuvent souvent rien; cette forme de taux de refus permet d'analyser l'efficacité d'un questionnaire ou l'effet du sujet d'une enquête sur le taux de participation des unités contactées. Par ailleurs, on utilise les équations (3.7) et (3.8) pour analyser le taux de refus comme un élément de la non-réponse parmi d'autres.

e) *Taux de réponse ou de non-réponse à une question*

À cause de la complexité du questionnaire, il arrive que des unités ne répondent pas à toutes les questions malgré leur bonne volonté (voir case 17). Par ailleurs, une question contraire ou personnelle ou encore une interruption de l'interview peut amener une unité à ne pas vouloir répondre à une question particulière (voir case 14). On peut donc mesurer par l'équation ci-dessous le taux de non-réponse global à la question "y" par rapport à l'ensemble des unités répondantes:

$$R_y = \frac{(\text{case 13})}{(\text{case 9}) + (\text{case 10})}$$

Si la question "y" ne s'adresse qu'à certaines unités, le taux de non-réponse à cette question peut être calculé uniquement en fonction des unités répondantes auxquelles s'adresse la question (unités admissibles). En conséquence, on peut définir toute une série de taux se rattachant aux questions (taux d'admissibilité, taux de réponse, taux de non-réponse) et qui sont comparables aux taux qui se rapportent aux questionnaires en remplaçant dans ces taux le nombre d'unités (admissibles/inadmissibles)/(acceptent de répondre/refusent de répondre, etc.) par

où l'unité i est admissible ou non à l'enquête selon que e_i est égale à 1 ou à 0. O'Neill, Groves et Cannell (1979) ont défini le rapport ci-dessus comme un taux d'intégralité. Dans la mesure où il est possible, comme en (3.3), de contrôler l'admissibilité de toutes les unités contactées, l'équation suivante (réelle ou estimée) définit probablement mieux les taux ci-dessus par rapport aux unités admissibles:

$$(3.4) \quad R^{(3)} = \frac{\sum_i t_i \delta_i e_i / \sum_i t_i e_i [\delta_i + (1 - \delta_i) r_i]}{[(\text{case } 9) + (\text{case } 10)] / [(\text{case } 9) + (\text{case } 10) + \text{case } 18]}.$$

Les taux (3.3) et (3.4) peuvent être utiles dans les enquêtes sur place ou par téléphone, où les cas de non-réponse peuvent être aussi bien des cas de non-contact que des cas de refus. Ils sont toutefois peu utiles dans les enquêtes par la poste à moins d'un suivi sur place ou par téléphone puisque dans la plupart de ces enquêtes les recenseurs ignorent forcément les motifs de réponse ou de non-réponse. Néanmoins, on reconnaît l'utilité de ces taux lorsqu'ils permettent d'évaluer dans quelle mesure une méthode de collecte de données incite les personnes désignées à répondre à l'enquête une fois qu'elles ont été contactées. Les cas de non-contact, contre lesquels les interviewers ne peuvent rien dans certaines enquêtes, sont entièrement exclus des taux. Le taux de réponse décrit en (3.4) a également été défini comme un taux d'intégralité par Klecka et Tuchfarber (1979), qui ont supposé, peut-être en manquant de réalisme, que toutes les unités qui avaient refusé de participer à l'enquête étaient admissibles. Le taux d'intégralité aurait alors été une estimation prudente de la capacité de la méthode de collecte à tirer des renseignements des unités admissibles. On peut par ailleurs supposer que le groupe d'unités ayant refusé de participer à l'enquête compte la même proportion d'unités admissibles que les groupes de répondants ou de non-répondants dont l'admissibilité a pu être contrôlée.

(c) *Taux de contact*

Selon Hauck (1974), un taux de contact est le pourcentage d'unités-échantillon qui ont été contactées; il est défini comme suit:

$$R^{(4)} = \frac{\text{Interviews achevées} + \text{cas de refus (unités contactées)}}{\text{Interviews achevées} + \text{cas de refus (unités contactées)} + \text{Interviews non-contactées et non-contactées}}$$

où l'on suppose que les unités non-contactées sont admissibles afin d'obtenir une estimation prudente du taux. Les cas de refus peuvent comprendre les interviews inachevées, qui équivalent essentiellement à un refus de répondre à certaines questions du questionnaire, comme l'indique la case 10 du tableau 1.

Le taux de contact est défini par l'expression algébrique suivante:

$$(3.5) \quad R^{(4)} = \frac{\sum_i t_i \delta_i e_i + \sum_i t_i (1 - \delta_i) r_i e_i}{\sum_i t_i \delta_i e_i + \sum_i t_i (1 - \delta_i) r_i e_i + \sum_i t_i (1 - \delta_i) r_i e_i}, \text{ ou } = \frac{(\text{case } 9) + (\text{case } 10) + (\text{case } 18)}{(\text{case } 9) + (\text{case } 10) + (\text{case } 18) + (\text{case } 19)},$$

$e_i = 1$ ou 0 si l'admissibilité a été vérifiée,

et si aucun contact n'a pu être fait

$e_i = 1$ selon la définition de Hauck,

ou $e_i = 0$, soit le taux d'admissibilité parmi les unités dont on a pu effectivement contrôler l'admissibilité.

les unités qui n'ont pu être contactées ou qui ont refusé de participer à l'enquête sont admissibles même si, en réalité, ces catégories de non-répondants ont souvent une proportion d'unités admissibles inférieure à celle des catégories de répondants et de non-répondants dont l'admissibilité a pu être contrôlée. L'hypothèse précédente permet de fixer une limite inférieure pour le taux de réponse et une limite supérieure pour le taux d'admissibilité. Par ailleurs, nous pouvons supposer que le groupe d'unités dont l'admissibilité ne peut être contrôlée a la même proportion d'unités admissibles que le groupe d'unités dont l'admissibilité peut être contrôlée. Cette hypothèse entraînerait vraisemblablement une légère surestimation du taux d'admissibilité et de quelques autres taux.

b) *Taux de réponse et taux d'intégralité*

i) Selon une des deux définitions établies par le U.S. Federal Committee on Statistical Methodology (1978), le taux de réponse est le pourcentage d'unités de l'échantillon admissibles pour lesquelles des données ont été recueillies. Le taux de réponse est donc défini comme suit:

$$R_{(1)} = \frac{\sum_i t_i e_i \delta_i}{\sum_i t_i e_i} \quad (3.2)$$

$$= [(case\ 9) + (case\ 10)] / (case\ 8).$$

C'est la formule la plus couramment utilisée dans la pratique puisqu'elle donne le pourcentage de l'échantillon pour lequel des données utiles ont été recueillies après élimination des unités inadmissibles. Le dénominateur, qui correspond au nombre d'unités admissibles, renferme toutes les catégories de non-répondants.

L'inverse du taux défini ci-dessus est souvent utilisé dans les cellules de correction comme facteur de correction de poids pour compenser la non-réponse, par exemple dans l'EPA du

Canada pour la correction de poids (voir Platek et Gray, 1985).

Le taux ci-dessus ou son complément, le taux de non-réponse, est souvent utilisé dans les évaluations administratives ou opérationnelles des organismes d'enquête. Il sert également à évaluer la capacité d'un intervieweur de contacter des répondants et d'obtenir leur collaboration pour leur faire communiquer des données utiles, par exemple à évaluer les taux de réponse ou de non-réponse par tâche d'interview. Le taux de non-réponse comprend aussi bien les cas de refus, qui peuvent être limités grâce à de bonnes relations publiques et à de la diplomatie, que les cas de non-contact, dont la cause peut échapper au contrôle de l'interviewer. Ainsi, dans la mesure du possible, les taux de non-réponse sont fréquemment ventiles selon le motif. Dans l'EPA, par exemple, le taux de réponse global se situe habituellement autour de 95%. Sur les 5 points de pourcentage qui correspondent au taux de non-réponse, environ 1 point est attribuable aux cas de refus.

Kviz (1977) a défini un taux semblable au précédent et l'a appelé taux d'intégralité; dans ce cas, le dénominateur comprend tout l'échantillon. Le taux d'intégralité offre probablement une estimation plus prudente du degré de qualité que le taux défini en (3.2) en ce sens que les unités inadmissibles, comme les logements inoccupés, sont comprises dans le dénominateur. En ce qui concerne l'EPA, par exemple, le taux de réponse ne serait plus de 95% mais bien de 85% si l'on appliquait la définition de Kviz.

ii) L'autre définition établie par le comité précité est le pourcentage de fois qu'un intervieweur obtient une interview d'une unité échantillonnée avec laquelle il est entré en contact; ce

taux est défini par l'équation suivante:

$$R_{(2)} = \frac{\sum_i t_i \delta_i}{\sum_i t_i [\delta_i + (1 - \delta_i) r_i]} \quad (3.3)$$

enquêtes par la poste, le fait qu'un questionnaire n'est pas retourné par l'unité choisie peut indiquer aussi bien un refus qu'une absence temporaire. Dans la perspective normale des études sur la non-réponse, on ne fait pas de distinction entre le fait que la personne désignée n'est pas à la maison et le fait qu'elle est absente pour quelque temps en ce qui concerne les enquêtes par la poste. Les motifs de non-réponse pour une enquête de ce genre doivent normalement être déterminés par un suivi sur place ou par téléphone en effectuant le plus souvent un sous-échantillonnage des non-répondants; certains d'entre eux peuvent finalement répondre au questionnaire tandis que d'autres continuent de compter parmi les non-répondants pour des raisons qu'il est possible de déterminer.

En règle générale, l'admissibilité d'une unité choisie est facilement vérifiable dans le cas d'une interview sur place; toutefois, si un interview ne peut contacter les unités choisies pour une enquête particulière, il peut ne pas pouvoir déceler les unités qui n'ont pas leur place dans cette enquête. En ce qui concerne les enquêtes par téléphone, le fait qu'il n'y a pas de réponse ou que la ligne est occupée peut empêcher l'interviewer de déterminer si l'unité est admissible ou non ou de définir le genre de non-réponse. On pourra également déduire qu'un répondant est inadmissible si l'on constate que le service téléphonique a été interrompu ou qu'une enquête de sélection a confirmé l'inadmissibilité de ce répondant. En ce qui a trait aux enquêtes par la poste, le fait qu'un questionnaire est retourné au point d'expédition pour diverses raisons (par exemple, adresse inexistante) peut être un indice de certaines formes d'inadmissibilité tandis que d'autres formes d'inadmissibilité ne peuvent être déterminées qu'au moyen d'un suivi sur place ou par téléphone.

3. DÉFINITIONS DE TAUX

La décomposition de l'échantillon de $n = \sum t_i$ unités illustrée dans le tableau 1 de la section précédente (unités admissibles ou inadmissibles, unités répondantes ou non-répondantes, refus ou non-refus, réponse ou non-réponse à une question, etc.) nous amène à définir ci-dessous de plusieurs genres de taux. Dans chaque cas, le numérateur est un sous-ensemble particulier du dénominateur. Dans la mesure du possible, les taux sont définis en fonction des formules de dénombrement indiquées dans l'organigramme du tableau 1.

(a) Taux d'admissibilité

Le taux d'admissibilité est défini par l'équation suivante:

$$\bar{e} = \sum_i t_i e_i / \sum_i t_i, \text{ (case 8)/(case 4).} \tag{3.1}$$

Wiseman et McDonald (1980) parlent dans ce cas d'un taux d'incidence mais ils n'appliquent ce terme qu'aux unités d'un échantillon qui ont effectivement accepté de participer à la phase de sélection d'une enquête, qui visait à déterminer leur admissibilité.

Le taux d'admissibilité défini en (3.1) indique dans quelle mesure le plan de sondage est efficace pour le prélèvement des unités admissibles dans une base de sondage dans le cas où il peut être difficile de contrôler l'admissibilité d'une unité sans un contact ou un examen rapide. Au stade de la présélection, le taux permet de savoir combien il y aura d'unités admissibles au moment de la collecte des données. Il peut donc servir à la conception d'une enquête si l'on dispose de données d'études antérieures sur l'admissibilité. Selon la nature et le déroulement de l'enquête, il peut être impossible de déterminer l'admissibilité d'unités qui n'ont pu être contactées ou même qui ont refusé de participer à l'enquête. Il y a deux façons de définir le taux d'admissibilité et le taux de réponse (que nous verrons plus loin) par rapport aux unités admissibles. Si nous voulons une évaluation prudente de la qualité des données et de la qualité de la méthode de collecte, nous pouvons supposer que toutes

Les $\sum t_i e_i \delta_i$ unités ayant répondu au questionnaire peuvent tout d'abord se diviser en deux groupes: $\sum t_i e_i \delta_i \Pi(\delta_{ip})$ unités ayant répondu à toutes les questions du questionnaire mais où il peut y avoir des erreurs de réponse (case 9) et $\sum t_i e_i \delta_i [1 - \Pi(\delta_{ip})]$ unités n'ayant pas répondu à toutes les questions du questionnaire (case 10). Dans les expressions ci-dessus, $\delta_{ip} = 1$ ou 0 selon que l'enquête a répondu ou non à la question "y". Dans la case 9, $\delta_{ip} = 1$ signifie que l'unité i a répondu à toutes les questions du questionnaire tandis que dans la case 10, $\delta_{ip} = 0$ signifie que l'unité i a négligé de répondre à au moins une question mais pas à toutes. Un certain nombre des $\sum t_i e_i \delta_i \delta_{ip}$ unités ayant répondu à la question "y" (case 12) viennent du groupe d'unités qui ont répondu à toutes les questions du questionnaire (case 9) tandis que les autres viennent du groupe d'unités qui ont omis de répondre à une ou à plus d'une question autre que la question "y". Les $t_i e_i \delta_i (1 - \delta_{ip})$ unités qui n'ont pas répondu à la question "y" (case 13) viennent du groupe d'unités qui ont omis de répondre à certaines questions (case 10), dont la question "y".

Les unités ayant répondu à la question "y" (case 12) peuvent être réparties en trois catégories: i) les unités pour lesquelles aucune erreur de réponse n'a été décelée, ii) celles pour lesquelles on a décelé une erreur de réponse à la question "y" et iii) celles pour lesquelles une erreur de réponse est passée inaperçue à la question "y"; ces groupes correspondent respectivement aux cases 15, 16A, et 16B.

Les $\sum t_i e_i \delta_i (1 - \delta_{ip})$ unités n'ayant pas répondu à la question "y" (case 13) sont toutes issues du groupe d'unités qui ont répondu au questionnaire; ce sont les unités pour lesquelles $\delta_i = 1$, $\delta_{ip} = 0$. Elles peuvent se diviser en deux groupes, c'est-à-dire i) celles qui ont refusé de répondre à la question "y" ou qui ont mis fins à l'interview avant d'être arrivées à cette question (case 14) et ii) celles qui n'ont pas pu répondre à la question "y" soit parce que ces unités ou l'interviewer ont mal compris le sens de la question ou parce qu'elles n'ont pas suivi l'ordre normal des questions.

Enfin, les unités n'ayant pas répondu au questionnaire (case 11) peuvent être classées parmi celles qui ont refusé de répondre au questionnaire (case 18) ou parmi celles qui n'ont pas répondu au questionnaire sans pour autant avoir refusé d'y répondre (case 19); ce dernier groupe compte surtout des unités qui n'ont pu être contactées parce qu'il n'y avait personne à la maison ou que la personne désignée pour répondre au questionnaire était absente pour quelque temps. On pose $r_i = 1$ pour les cas de refus et $r_i = 0$ pour les autres motifs de non-réponse. Parmi ces derniers, on relève surtout les deux suivants: "personne à la maison" et "absent du foyer pour quelque temps".

Pour pouvoir dénombrer les répondants et les non-répondants selon la catégorie et le motif, il faut enregistrer soigneusement chaque unité échantillonnée. Une telle mesure est essentielle si nous voulons éviter qu'un échantillon aléatoire dégénère en un échantillon par la méthode des quotas à cause, par exemple, d'un traitement improvisé de la non-réponse, qui se traduirait notamment par une substitution arbitraire de nouvelles unités aux non-répondants. En ce qui concerne les échantillons par la méthode des quotas, il est parfois difficile ou impossible de distinguer les unités substituées de celles qui ont été choisies à l'origine lorsque les recenseurs essaient de compléter l'échantillon en se tournant vers des répondants plus coopératifs au lieu d'insister auprès des non-répondants.

Même lorsque nous avons des échantillons aléatoires dont les unités ont été soigneusement identifiées et prélevées selon le plan, il est parfois difficile de déterminer le motif exact de la non-réponse en ce qui concerne les unités qui n'ont pu être contactées. Le cas le plus simple est celui des interviews sur place. Toutefois, même lorsqu'il s'agit d'une interview sur place, il peut être difficile de faire la distinction entre le fait qu'il n'y a personne à la maison et le fait que la personne désignée est absente du foyer pour quelque temps ou encore entre un cas de refus et un cas de non-contact lorsqu'il y a manifestement quelqu'un dans la maison mais qu'on refuse d'ouvrir. En ce qui a trait aux interviews par téléphone, le fait qu'il n'y ait pas de réponse ou que la ligne soit occupée ne signifie pas nécessairement que l'unité choisie n'a pas été contactée; en revanche, il est facile de constater un cas de refus par téléphone. En ce qui concerne les

La non-réponse à une question a généralement trait aux questionnaires proprement dits; dans ce cas, les enquêtes n'ont pas répondu à toutes les questions du questionnaire. Cependant, si une unité ne répond pas du tout au questionnaire, il va de soi qu'elle ne répond à aucune des questions. Par conséquent, la non-réponse à un questionnaire et la non-réponse à une question sont des éléments distincts qui doivent être considérés séparément.

Les taux de réponse peuvent se rapporter à tout l'échantillon ou à une partie de celui-ci (par exemple à des secteurs liés au plan de sondage) ou ils peuvent s'appliquer à des secteurs administratifs, comme la zone d'affectation d'un intervieweur ou un groupe de zones d'affectation sous la direction d'un superviseur ou d'un bureau régional.

2. ÉLÉMENTS DE LA RÉPONSE ET DE LA NON-RÉPONSE

Pour définir divers taux de réponse et analyser leurs applications, il faut répartir la population cible du sondage ou du recensement en diverses composantes selon le genre de réponse ou de non-réponse. C'est exactement ce que montre le tableau 1, où figurent les principaux éléments de l'enquête qui entreront dans la définition des taux. Une fois définie la population cible (case 1), on détermine une base de sondage de N unités (case 2).

Comme un sur ou sous-dénombrement d'unités est toujours possible, il convient de souligner que la base de sondage peut ne pas correspondre parfaitement à la population cible. Comme le sous-dénombrement est plus fréquent que le surdénombrement dans la pratique, la population cible réelle comprend habituellement plus que N unités.

En vue de réaliser l'enquête, on choisit une méthode de collecte des données (case 3) et on détermine un plan de sondage approprié. Qu'il s'agisse d'un sondage ou d'un recensement, nous avons les éléments suivants:

$$n = \sum t_i \text{ unités sont choisies, où:}$$

$$t_i = 1 \text{ ou } 0, \text{ selon que l'unité } i \text{ est choisie ou non,}$$

$$\sum = \text{somme pour les } N \text{ unités comprises dans la base de sondage.}$$

Dans une base de sondage, il arrive souvent qu'on ignore la valeur exacte de N et qu'on ne puisse l'estimer qu'au moyen de l'échantillon. C'est souvent ce qui se produit dans les échantillonnages aléatoires à plusieurs degrés qui comportent au début un sondage aréolaire. Des n unités de l'échantillon, $\sum t_i e_i$ sont admissibles (case 8) et $\sum t_i (1 - e_i)$ sont inadmissibles (case 5) aux fins de l'enquête, où

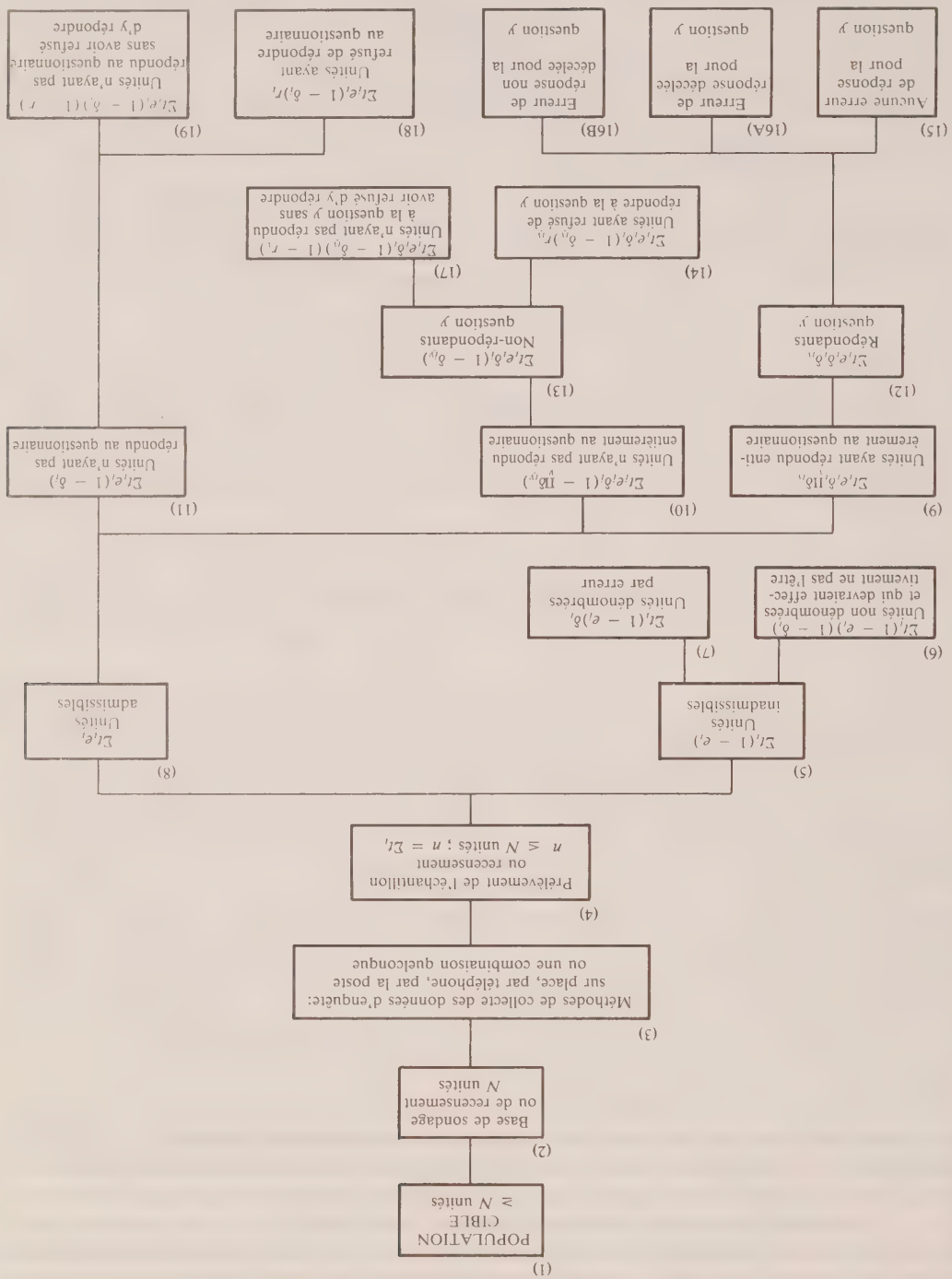
$$e_i = 1 \text{ ou } 0, \text{ selon que l'unité } i \text{ est admissible ou non.}$$

Il arrive que l'on ne puisse déterminer si une unité est admissible ou non tant et aussi longtemps qu'elle n'a pas été contactée, tandis que dans d'autres cas, l'admissibilité est facilement établie par simple constatation de fait (par exemple logement inoccupé, dans une enquête sur les ménages).

Les $\sum t_i (1 - e_i)$ unités inadmissibles de la case 5 peuvent être réparties en deux groupes: $\sum t_i (1 - e_i)$ unités qui n'ont pas été interviewées, comme prévu (case 6), et $\sum t_i (1 - e_i)$ unités qui ont été interviewées par erreur (case 7). Il est souhaitable que le nombre d'unités dans cette dernière case soit nul ou, du moins, très faible. Si toutefois on décelait de telles unités, il faudrait les soustraire de l'échantillon. Dans note exposé, $\delta_i = 1$ ou 0 selon que l'unité i a répondu ou non au questionnaire.

Les $\sum t_i e_i$ unités admissibles (case 8) peuvent, elles aussi, être réparties en deux groupes principaux: $\sum t_i e_i \delta_i$ unités ayant répondu au questionnaire (cases 9 et 10) et $\sum t_i e_i (1 - \delta_i)$ unités non-répondantes (case 11), c'est-à-dire que celles-ci n'ont fourni aucune donnée utilisable et que l'on connaît très peu de choses à leur sujet hormis, peut-être, leur situation géographique.

Tableau 1
Éléments de la réponse et de la non-réponse



$e_i = 1,0$ (selon que l'unité est admissible ou non)
 $\delta_i = 1,0$ (selon que l'unité a répondu ou non à la question y)
 $r_i = 1,0$ (suivant que l'unité a refusé ou non de répondre)
Lorsque $r_i = 0$, et $\delta_i = 0$, il s'agit surtout d'unités dont les membres n'étaient pas au foyer au moment de l'enquête ou du recensement ou s'étaient temporairement absents.

l'importance du biais dû à la non-réponse par le rapport de ce biais au coefficient de la variation d'échantillonnage, alors l'importance de ce biais est proportionnelle au produit de la racine carrée de la taille de l'échantillon répondant par le taux de non-réponse.

En termes plus pratiques, le taux de réponse (ou de non-réponse) peut refléter les problèmes de fonctionnement qui existent dans un recensement ou une enquête et fournir un indice de fiabilité des données d'enquête. Toutefois, divers genres de taux de réponse (ou de non-réponse) sont utilisés à ces fins, selon qu'une unité désignée a pu être contactée ou non. On peut alors distinguer les cas où il y a eu un contact d'établi de ceux où il n'y en a pas eu. Lorsqu'un cas de non-réponse s'explique par le fait qu'il n'y avait personne à la maison ou que la personne choisie pour l'enquête était temporairement absente, il s'agit en fait d'un cas où il n'y a eu aucun contact d'établi et, par le fait même, le problème est surtout de nature opérationnelle. On parle d'un véritable cas de non-réponse lorsque l'unité choisie a été contactée mais n'a pas voulu répondre à l'enquête ou n'a pas fourni de réponse acceptable. En traitant un échantillon d'enquête, un intervieweur peut constater la présence d'unités qui, normalement, ne devraient pas faire partie de cet échantillon (unités inadmissibles). Par ailleurs, certaines unités ne répondront que partiellement au questionnaire tandis que d'autres y répondront entièrement. Chacun de ces cas peut être représenté par un taux, par exemple un taux d'admissibilité, un taux de réponse à une question, un taux d'intégralité, etc. La distinction entre les cas réels de non-réponse et les autres cas qui influent sur le taux de non-réponse peut donner lieu à diverses interprétations.

Il est particulièrement difficile d'interpréter les taux de réponse (ou de non-réponse) lorsqu'il s'agit de plans de sondage complexes puisque le taux de non-réponse peut être plus élevé dans un secteur ou une catégorie donnés que dans un autre. Or presque tous les statisticiens utilisent les taux de réponse comme indices approximatifs de la qualité des données. C'est pourquoi, à chaque enquête, on cherche toujours à recueillir des données sur la non-réponse et à évaluer l'ampleur du phénomène. Cependant, l'évaluation des résultats d'une enquête ne peut reposer sur des bases solides que si elle est faite à partir des mesures d'erreurs systématiques, de variances et d'erreurs quadratiques moyennes correspondantes tirées de toutes les sources d'erreur (erreurs d'échantillonnage et d'observation).

Depuis quelques années, le taux de non-réponse s'est accru dans beaucoup d'enquêtes au Canada et à l'étranger. Il est donc plus que jamais nécessaire de contrôler les taux de non-réponse, d'établir des comparaisons entre les enquêtes, les pays et les organismes d'enquête et d'établir une base de comparaison acceptable. Des efforts ont été faits pour uniformiser la définition du taux de réponse et de son complètement, le taux de non-réponse; voir à ce sujet Kviz (1977) et Cannell (1978). Wiseman et McDonald (1980) font état de définitions disparates des taux de réponse dans des enquêtes par téléphone.

Le phénomène de non-réponse dans les enquêtes entraîne l'emploi d'un vocabulaire peu uniforme. Des expressions comme taux d'intégralité, taux de contact et taux de sous-dénombrément ont reçu des acceptions diverses dans des rapports et des articles sur la collecte de données. Bien qu'il soit facile de distinguer ces expressions à l'intérieur d'un même rapport, ils peuvent porter à confusion et faire l'objet d'interprétations contradictoires lorsqu'on les retrouve dans différents rapports.

Avant d'analyser le problème de la non-réponse, il est nécessaire de faire la distinction entre les taux de non-réponse à un questionnaire et les taux de non-réponse à une question. Les premiers se rattachent habituellement au niveau auquel les données d'enquêtes sont recueillies au moment du premier contact. Il peut s'agir, par exemple, d'un logement, d'une personne, d'un magasin ou d'un établissement. Dans le cas d'un échantillonnage à plusieurs degrés, toutefois, il se peut qu'aucune des unités d'une grappe, voire d'une unité primaire d'échantillonnage (u.p.é.), ne réponde au questionnaire, de sorte qu'un taux de non-réponse au questionnaire pourrait s'appliquer aussi bien à une grappe ou une u.p.é. donnée qu'à un logement ou à une personne.

Sur les définitions des taux de réponse

R. PLATEK et G.B. GRAY¹

RÉSUMÉ

Les auteurs définissent et analysent divers genres de réponse ou de non-réponse ainsi que les indices, notamment des taux, qui servent à mesurer ces phénomènes; ils étudient également les effets de ces phénomènes sur les méthodes d'estimation et les procédés administratifs. Le problème de la non-réponse donne naissance à des expressions variées comme taux d'intégralité, taux d'admissibilité, taux de contact et taux de refus, et plusieurs de ces termes peuvent avoir des définitions variées. Il y a aussi les taux de non-réponse à une question et les taux de réponse par caractéristique. Suivant l'usage qu'on en fait, les taux peuvent être pondérés ou non.

MOTS CLÉS: Admissibilité; intégralité; contact; refus; taux de réponse.

1. INTRODUCTION

Les données d'un recensement ou d'une enquête par sondage peuvent être recueillies au moyen d'une interview sur place ou d'une interview téléphonique ou encore par la poste. Il arrive parfois que l'on ne puisse obtenir de réponse de certaines unités soit parce que la personne désignée est temporairement absente du foyer ou qu'elle est partie en vacances, soit parce que l'établissement visé (dans le cas d'enquêtes commerciales) est fermé, soit parce que le répondant a refusé de participer à l'enquête ou que l'unité est inoccupée ou démolie, etc. Il peut arriver aussi que des unités ne répondent que partiellement au questionnaire; par exemple, il se peut que seulement quelques membres d'un ménage répondent au questionnaire ou que les unités visées ne répondent pas à toutes les questions. Par ailleurs, des unités peuvent fournir des réponses inexacts ou imprécises.

Ainsi, quelles que soient l'enquête et la méthode de collecte, il y aura toujours des données manquantes à cause de la non-réponse. Le taux de non-réponse a toujours été considéré comme un indice important de la qualité des données puisque ce phénomène introduit vraisemblablement une erreur systématique dans les estimations et qu'il contribue à accroître la variance d'échantillonnage à cause de la diminution de la taille de l'échantillon initial. La relation entre la variance d'échantillonnage et le taux de non-réponse est assez claire. Toutefois, même si elle peut être plus importante, la relation entre l'erreur systématique et le taux de non-réponse est moins évidente puisqu'elle dépend à la fois de l'ampleur de la non-réponse et des différences de caractéristiques entre les répondants et les non-répondants. On pourrait supposer que l'erreur systématique due à la non-réponse est proportionnelle au taux de non-réponse. Pour un taux de réponse donné, l'erreur systématique en pourcentage serait alors indépendante de la taille de l'échantillon. Or, la variance d'échantillonnage est influencée par la taille de l'échantillon et inversement proportionnelle à la taille de l'échantillon répondant. Par conséquent, le rapport entre le biais dû à la non-réponse et les erreurs d'échantillonnage peut n'être pas aussi élevé pour les petits échantillons que pour les grands. Lorsqu'il y a un biais dû à la non-réponse, l'intervalle de confiance vraisemblable peut inclure la valeur recherchée s'il s'agit d'un petit échantillon mais risque de ne pas l'inclure s'il s'agit d'un grand échantillon. Si nous mesurons

¹ R. Platek, ancien directeur de la Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada; G.B. Gray, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario K1A 0T6.

- OH, H.T., SCHEUREN, F., et NISSELTSON, H. (1980). Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 416-420.
- PALMER, S. (1967). On the character and influence of nonresponse in the Current Population Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 73-80.
- PALMER, S., et JONES, C. (1966). A look at alternate imputation procedures for CPS noninterviews. Washington, D.C.: U.S. Bureau of the Census document interne.
- POLITZ, A., et SIMMONS, W. (1949). I. An attempt to get the 'not at homes' into the sample without callbacks. II. Further theoretical considerations regarding the plan for eliminating callbacks. *Journal of the American Statistical Association*, 44, 9-31.
- POLITZ, A., et SIMMONS, W. (1950). Note on an attempt to get the 'not at homes' into the sample without callbacks. *Journal of the American Statistical Association*, 45, 136-137.
- RUBIN, D.B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-34.
- RUBIN, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. *Bulletin of the International Statistical Institute*, 48(2), 517-532.
- RUBIN, D.B., et SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- SANDE, G. (1979). Numerical edit and imputation. Article présenté à l'International Association for Statistical Computing, 42nd Session of the International Statistical Institute.
- SANDE, I.G. (1983). Hot-deck imputation procedures. Dans *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (éd. W.G. Madow et I. Olkin), New York: Academic Press, 339-349.
- SANTOS, R.L. (1981). Effects of imputation on regression coefficients. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 140-145.
- THOMSEN, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidsskrift*, 4, 278-283.
- THOMSEN, I., et SIRING, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. Dans *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, (éd. W.G. Madow et I. Olkin), New York: Academic Press, 25-29.
- VACEK, P.M., et ASHIKAGA, T. (1980). An examination of the nearest neighbor rule for imputing missing values. *Proceedings of the Statistical Computing Section, American Statistical Association*, 326-331.
- WELNIAR, E.J., et CODER, J.F. (1980). A measure of the bias in the March CPS earnings imputation system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 421-425.

- COLLEDGE, M.J., JOHNSON, J.H., PARE, R., et SANDE, I.G. (1978). Large scale imputation of survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 431-436.
- COX, B.G., et COHEN, S.B. (1985). *Methodological Issues for Health Care Surveys*. New York: Marcel Dekker.
- DAVID, M., LITTLE, R.J.A., SAMUEL, M.E., et TRIEST, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.
- DREW, J.H., et FULLER, W.A. (1980). Modelling nonresponse in surveys with callbacks. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-642.
- DREW, J.H., et FULLER, W.A. (1981). Nonresponse in complex multiphase surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 623-628.
- FORD, B.L. (1983). An overview of hot-deck procedures. Dans *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (éd. W.G. Madow, I. Olkin et D.B. Rubin), New York: Academic Press, 185-207.
- GREENLEES, W.S., REECE, J.S., et ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- HERZOG, T.N., et RUBIN, D.B. (1983). Using multiple imputation to handle nonresponse in sample surveys. Dans *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (éd. W.G. Madow, I. Olkin et D.B. Rubin), New York: Academic Press, 209-245.
- KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor: Survey Research Center, University of Michigan.
- KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, en voie de rédaction.
- KALTON, G., et KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22-31.
- KALTON, G., et KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics - Theory and Methods*, 13(16), 1919-1939.
- KISH, L. (1965). *Survey Sampling*. New York: Wiley.
- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, Ser. A*, 139, 80-95.
- LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- LITTLE, R.J.A. (1986a). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- LITTLE, R.J.A. (1986b). Missing data in Census Bureau surveys. *Proceedings of the Second Annual Census Bureau Research Conference*, 442-454.
- LITTLE, R.J.A., et DAVID, M.H. (1983). Weighting adjustments for non-response in panel surveys. Document de travail. Washington, D.C.: U.S. Bureau of the Census.
- OH, H.T., et SCHEUREN, F. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.
- OH, H.T., et SCHEUREN, F. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 723-728.
- OH, H.T., et SCHEUREN, F. (1980). Estimating the variance impact of missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 408-415.
- OH, H.T., et SCHEUREN, F. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete data in Sample Surveys, Volume 2, Theory and Bibliographies*, (éd. W.G. Madow, I. Olkin et D.B. Rubin), New York: Academic Press, 143-184.

Une hypothèse se trouvant à la base de toutes les procédures examinées ici est celle selon laquelle une fois que les variables auxiliaires ont été prises en compte, les valeurs manquantes ne sont au hasard. Ainsi, par exemple, on suppose que les non-répondants sont comme les répondants à l'intérieur des classes de pondération et d'imputation. Cette hypothèse peut être évitée en utilisant des modèles de censure stochastiques, comme l'ont fait Greenlees *et al.* (1982) en imputant les salaires et traitements de la Current Population Survey. Cependant, comme Little (1986b) le fait remarquer, ces modèles sont très sensibles aux hypothèses de distribution retenues.

Une autre méthode pour le traitement des données d'enquête manquantes consisterait à laisser les valeurs manquantes dans l'ensemble de données et laisser l'analyste incorporer des modèles de données manquantes appropriés dans l'analyse (Little 1982). Cette méthode a beaucoup de caractéristiques intéressantes, mais les ressources en personnel et calculs nécessaires à sa mise en oeuvre l'empêchent de l'utiliser comme stratégie générale. Cette approche semble plutôt mieux convenir à un petit éventail d'analyses spéciales. Afin de permettre à l'analyste d'adopter cette approche, il est essentiel que toutes les valeurs imputées soient indiquées afin de signaler lesquelles ne sont pas des réponses réelles, de sorte qu'il soit possible de les laisser de côté dans l'analyse.

Enfin, nous devons noter que toutes les méthodes de traitement des données manquantes d'enquête doivent dépendre d'hypothèses non testables. Si les hypothèses sont gravement erronées, les analystes peuvent donner des conclusions erronées. La seule façon sûre d'éviter d'importants biais de non-réponse dans les estimations d'enquête est de limiter la quantité de données manquantes.

BIBLIOGRAPHIE

- BAILLAR III, J.C., et BAILLAR, B.A. (1978). Comparison of two procedures for imputing missing survey values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 462-467.
- BAILLAR, B.A., et BAILLAR III, J.C. (1983). Comparison of the biases of the hot-deck imputation procedure with an "equal-weights" imputation procedure. Dans *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*. (éd. W.G. Madow et I. Olkin), New York: Academic Press, 299-311.
- BAILLAR, B.A., BAILEY, L., et CORBY, C.A. (1978). A comparison of some adjustment and weighting procedures for survey data. Dans *Survey Sampling et Measurement*, (éd. N.K. Namboodiri), New York: Academic Press, 175-198.
- BARTHOLOMEW, D.J. (1961). A method of allowing for 'not at home' bias in sample surveys. *Applied Statistics*, 10, 52-59.
- BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analyses*. Cambridge, Mass: The MIT Press.
- BROOKS, C.A., et BAILLAR, B.A. (1978). *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Policy Working Paper 3. U.S. Department of Commerce, Washington, D.C.: U.S. Government Printing Office.
- CHAPMAN, D.W., BAILEY, L., et KASPRZYK, D. (1986). Nonresponse adjustment procedures at the U.S. Census Bureau. *Survey Methodology*, en voie de rédaction.
- CODER, J. (1978). Income data collection and processing from the March Income Supplement to the Current Population Survey. *The Survey of Income and Program Participation Proceedings of the Workshop on Data Processing*, February 23-24, 1978, (éd. D. Kasprzyk), Chapter II. Washington, D.C.: U.S. Department of Health, Education and Welfare.

répétitions, m . Pour cette raison, on peut préférer une valeur de m peu élevée, comme par exemple $m = 2$. Une telle valeur de m peu élevée peut cependant se traduire par un niveau de précision bas pour l'estimateur de la variance. Même avec un m peu élevé, on peut se demander si la méthode de l'imputation multiple est plausible pour les analyses de routine. Elle serait mieux utilisée dans des études spéciales comme celles décrites par Herzog et Rubin (1983).

En plus de fournir des erreurs types appropriées, un autre avantage des imputations multiples résultant de la même procédure d'imputation est qu'elle réduit la perte de précision dans les estimations d'enquête résultant de la sélection aléatoire de répondants comme donneurs de valeurs imputées (voir section 3.1). Cette perte se trouve réduite dans le cas des imputations multiples par une mise en moyenne des répétitions. Un petit nombre de répétitions convient bien à cette fin. Comme on l'a dit plus tôt, Kalton et Kish (1984) décrivent d'autres moyens de choisir l'échantillon de répondants pour y arriver.

Une autre application potentielle importante des imputations multiples est la production d'imputations pour les différentes répétitions grâce à des procédures d'imputation différentes, retenant des hypothèses différentes à propos des non-répondants. Supposons par exemple que les taux horaires de rémunération doivent être imputés pour certains salariés de l'échantillon. Une procédure qui pourrait être utilisée est la méthode de l'imputation aléatoire intra-classe, qui repose sur l'hypothèse que les non-répondants manquent au hasard à l'intérieur des classes. Si l'on estime que les non-répondants pourraient dans la réalité provenir d'avant-tage des salariés touchant des taux plus élevés de rémunération dans chaque classe, une simple modification de la méthode aléatoire intra-classe consisterait à imputer des valeurs qui dépasseraient de 50 cents les valeurs des donneurs, par exemple. D'autres procédures d'imputation, utilisant par exemple des classes d'imputation différentes, pourraient être également essayées. La comparaison des estimations d'enquête obtenues des ensembles de données auxquels on a appliqué les différentes procédures d'imputation fournit alors une indication importante de la sensibilité des estimations aux valeurs imputées. Si les estimations se révèlent être très semblables, on peut les accepter avec plus de confiance; si elles diffèrent sensiblement, il faut les traiter avec beaucoup de prudence.

4. CONCLUSION

On a présenté la pondération et l'imputation comme deux méthodes distinctes pour le traitement des données d'enquête manquantes, mais en fait il existe une relation étroite entre elles. À titre d'illustration, on peut considérer une méthode d'imputation qui attribue les valeurs des répondants aux non-répondants. Dans les analyses univariées, ce procédé équivaut à l'abandon des enregistrés des non-répondants et à l'addition des poids des non-répondants à ceux des répondants (Kalton 1986).

Les différences entre la pondération et l'imputation ressortent lorsqu'on examine la nature multivariée des données d'enquête. Il est possible d'imputer les réponses d'un non-répondant intégral en prenant toutes les réponses d'un donneur unique; cependant, la pondération est généralement plus simple dans ce cas et elle permet d'éviter la perte de précision résultant de l'échantillonnage de répondants devant servir comme donneurs. Il n'est pas pratique d'utiliser la pondération pour régler la non-réponse à une question, puisque cela se traduirait par des ensembles différents de poids pour chaque question; cette façon d'agir entraînerait de graves difficultés dans le cas des totalisations recoupées et d'autres analyses des relations entre variables. La pondération est un ajustement global unique qui essaie de compenser les réponses manquantes pour toutes les questions simultanément. L'imputation, par contre, traite chaque question séparément. Cette différence se répercute sur la façon dont les données auxiliaires sont utilisées. Lors de la constitution des classes de pondération, l'objectif et de déterminer les classes dont le taux de réponse diffèrent. Le choix des variables auxiliaires utilisées dans l'imputation, cependant, est déterminé d'abord en fonction de leurs possibilités de prédire les réponses manquantes.

et imputés ayant le caractère est par conséquent $(80 + 60 + 60 + 40)/400 = 0,6$, ou 60%. Pour les femmes, la proportion correspondante est $(40 + 60 + 20 + 40)/400 = 0,4$, ou 40%. La différence entre ces deux pourcentages est de 20%.

Si l'on avait également pris le sexe en compte dans la formation des classes d'imputation, les pourcentages des hommes et des femmes possédant le caractère auraient été de 70% et de 30%, soit une différence de 40%. La non-inclusion du sexe comme variable auxiliaire dans l'imputation s'est traduite ainsi par une réduction appréciable de la mesure de la relation entre le sexe et la présence du caractère.

3.4 Imputations multiples

Idéalement, l'analyste qui utilise un ensemble de données avec des valeurs imputées devrait être en mesure d'obtenir des résultats valables pour toute analyse en utilisant pour cela des techniques normalisées pour les données complètes. Cependant, comme on l'a remarqué dans la section précédente, l'imputation peut déformer les mesures des relations entre variables. Elle peut également déformer l'estimation de l'erreur type.

Toutes les méthodes d'imputation à l'exception de l'imputation par déduction fabriquent dans une certaine mesure les données. L'ampleur de cette fabrication dépend de la qualité de la prédiction des valeurs manquantes par le modèle d'imputation. Si ce dernier n'explique qu'une petite partie de la variance dans la variable entre les répondants, la quantité de fabrication dans chaque valeur imputée sera probablement appréciable. Si le modèle d'imputation explique une proportion élevée de la variance des répondants, cette quantité sera probablement moins importante. Cependant, il ne faut pas perdre de vue que l'ajustement du modèle d'imputation pour les répondants n'est pas nécessairement une bonne mesure de l'ajustement pour les non-répondants.

Les erreurs types calculées de la manière habituelle à partir d'un ensemble de données contenant des valeurs imputées seront généralement sous-estimées en raison du degré de fabrication contenu dans les valeurs imputées. Rubin (1978, 1979) a suggéré l'emploi de la méthode des imputations multiples pour obtenir des inférences valides à partir d'ensembles de données contenant des valeurs imputées (voir également Herzog et Rubin 1983, et Rubin et Schenker 1986). Lorsqu'on utilise les imputations multiples aux fins de l'estimation de l'erreur type, la construction de l'ensemble complet des données par imputation des réponses manquantes s'effectue en plusieurs fois (m par exemple) en utilisant la même procédure de l'imputation. Les estimations de l'échantillon z_i ($i = 1, 2, \dots, m$) du paramètre de la population étudiée Z sont calculées à partir de chacun des ensembles de données répétées, et on calcule leur moyenne \bar{z} . Un estimateur de la variance pour \bar{z} est ensuite obtenu par la formule $V = W + [(m + 1)/m]B$, où W est la moyenne de la variance intra-répétition de \bar{z} et $B = \sum (z_i - \bar{z})^2 / (m - 1)$, les étendues des intervalles de confiance pour Z basées sur V sont toujours surestimées, le degré de surestimation augmentant avec le niveau de non-réponse. Cette surestimation des niveaux de confiance peut être corrigée en modifiant la procédure d'imputation, comme le décrivent Rubin et Schenker (1986). Ils considèrent la méthode d'imputation globale aléatoire, et l'une de leurs modifications prend en compte l'incertitude entourant la moyenne et la variance de la population de la façon suivante. Avec la méthode d'imputation globale aléatoire normale, les moyenne et variance attendues conditionnelles des valeurs imputées sont la moyenne et la variance des répondants de l'échantillon. Avec la modification, la moyenne et la variance attendues des valeurs imputées d'une répétition sont tirées au hasard des distributions appropriées. Les valeurs imputées sont donc une sélection aléatoire de valeurs de répondants, modifiées pour la moyenne et la variance choisies au hasard. Lorsque l'on estime la moyenne de la population, l'effet du changement de la variance intra-répétitions est d'accroître la composante de la variance attendue de la variance intra-répétitions dans V . Ceci se traduit par une meilleure délimitation des intervalles de confiance obtenus.

Un problème important posé par l'utilisation d'imputation multiples est la quantité d'analyses informatiques supplémentaires nécessaires, qui augmente avec le nombre de

où $S_{xy,z}^h = \Sigma W^h S_{xy}^h$ est la covariance intra-classe moyenne pour les classes formées par les variables auxiliaires z , S_{xy}^h est la covariance à l'intérieur de la classe h , et W^h est la proportion de la population de la classe h . Avec une imputation par régression prédite ou une imputation par régression avec un résidu aléatoire, toutes deux avec une variable auxiliaire unique z , le biais relatif est approximativement $-[M(1 - (\rho_{xz}\rho_{yz}/\rho_{xy}))]$, où ρ^{uv} est la corrélation entre u et v .

La caractéristique inquiétante de ce résultat est que, à moins que M ne soit petit, s_{xy} calculé avec des valeurs imputées en vertu de l'une de toutes ces méthodes d'imputation peut avoir un biais appréciable, même dans le cas du modèle avec données manquantes au hasard. Les estimations s_{xy} calculées avec des valeurs imputées obtenues grâce aux méthodes d'imputation de classe et par régression sont sans biais seulement si la covariance partielle $S_{xy,z}$ est zéro. En général, il n'y a pas de raison de supposer les yeux fermés que $S_{xy,z}$ est 0. Toutefois, il y a un important cas lorsque $S_{xy,z} = 0$. Ceci se produit lorsque $x = z$, c'est-à-dire lorsque x est utilisé comme variable auxiliaire dans la procédure d'imputation. Dans ce cas, la covariance de l'échantillon est sans biais selon le modèle de données manquantes au hasard. Ce résultat permet de croire que si la relation entre x et y doit former une partie importante de l'analyse d'enquête, x devrait être utilisé comme variable auxiliaire dans l'imputation des données manquantes y .

La théorie ci-dessus suppose qu'il a des données manquantes seulement pour la variable y . Dans la pratique, la variable x sera également souvent incomplète. Dans ce cas, la covariance d'échantillon peut être réduite en raison des imputations pour les deux variables. Une situation particulière se présente lorsque x et y manquent dans un enregistrement. Si les deux valeurs font l'objet de deux imputations distinctes, la covariance est réduite, mais si elles sont imputées conjointement, en utilisant le même répondant comme le donneur des deux valeurs, la structure de covariance subsiste. Ceci permet de croire que lorsqu'un enregistrement a plusieurs valeurs connexes manquantes, elles doivent être prises du même donneur. Coder (1978) décrit l'utilisation de l'imputation conjointe du même donneur dans l'enquête supplémentaire sur le revenu de mars de la Current Population Survey.

A titre d'illustration de la façon dont les arguments ci-dessus à propos de l'atténuation des covariances s'appliquent à d'autres formes de relation, nous allons donner un exemple numérique simple de l'effet de l'imputation sur la différence entre deux proportions. Supposons que la variable qui nous intéresse soit le fait qu'une personne ait un caractère particulier ou non, et supposons que la moitié des répondants ne répondent pas à la question. Les réponses manquantes sont imputées par une méthode aléatoire d'imputation intra-classe qui utilise deux classes, A et B . L'objectif est maintenant de comparer les pourcentages des hommes et des femmes possédant ce caractère. Les données sont présentées au tableau 1. Comme 60% du total des répondants de la classe A ont le caractère, 60 des 100 hommes et 60 des 100 femmes non-répondants de cette classe seront imputés comme ayant le caractère. De même, dans la classe B , 40% du total des répondants ont le caractère, et 40 hommes et 40 femmes non-répondants seront imputés comme l'ayant. La proportion des hommes réels

Tableau 1
Nombre de répondants possédant le caractère et nombre de personnes échantillonnées selon la classe, le sexe et le type de réponse.

Classe A			Classe B		
H	F	Total	H	F	Total
80	40	120	60	20	80
100	100	200	100	100	200
100	100	200	100	100	200
Non-répondants			Total, échantillon		
200	200	400	200	200	400

une modélisation minutieuse, on court le risque sérieux d'avoir de mauvaises imputations, bien que, comme on l'a mentionné plus tôt, ce risque peut être réduit par l'affectation de résidus aléatoires de répondants "proches".

Si l'imputation par régression attribue au résidu d'un répondant exactement les mêmes valeurs des variables auxiliaires, la valeur imputée sera nécessairement une valeur acceptable. Cependant, s'il existe une différence, même minime, entre les valeurs du répondant et du non-répondant pour les variables auxiliaires, la valeur imputée peut ne pas être acceptable. Une variante de l'imputation par régression qui évite ce problème, appelée appariement de moyennes prédictives est décrite par Little (1986b) (Little attribue cette méthode à Rubin). Selon cette méthode, le non-répondant est apparié au répondant qui possède la valeur prédite la plus proche. Alors, au lieu d'ajouter le résidu du répondant à la valeur prédite du non-répondant, on attribue à ce dernier la valeur du répondant. La méthode devient une méthode hot-deck semblable à l'appariement de la fonction de distance.

Le choix entre les méthodes de classe d'imputation et d'imputation par régression devrait dépendre en partie des efforts consacrés à la mise au point des modèles de régression. À moins que des ressources appropriées ne soient consacrées à la mise au point d'un modèle de régression, la méthode de classe d'imputation pourrait être plus sûre. Le choix doit également en partie dépendre de la taille de l'échantillon. Lorsque l'échantillon est important, les méthodes hot-deck vont probablement utiliser suffisamment de classes pour bénéficier de toutes les principales variables indépendantes; cependant, dans le cas des petits échantillons, ceci ne pourrait pas être le cas, et les méthodes de régression pourraient présenter un plus grand potentiel. David *et al.* (1986) décrivent une intéressante étude qui compare les modèles de régression pour l'imputation des salaires et traitements dans la U.S. Current Population Survey aux imputations hot-deck hiérarchiques. En dépit des efforts importants consacrés à la mise au point de modèles de régression, les imputations hot-deck ne devaient pas se révéler inférieures dans cet important échantillon.

3.3 Effet de l'imputation sur les relations

Bien que la plus grande partie de la bibliographie consacrée à l'imputation traite de son effet sur les statistiques univariées telles que les moyennes et les distributions, une importante partie de l'analyse d'enquêtes se rapporte aux relations bivariées et multivariées. Dans ce cas, l'analyse des relations peut être considérée en termes généraux de façon à regrouper les totalisations recoupées, la corrélation ou l'analyse de régression, la comparaison des moyennes ou des proportions intra-classe et toute autre analyse faisant intervenir deux ou plusieurs variables. Comme on le montrera ci-dessous, l'imputation peut avoir des effets nocifs sur toutes les analyses de relations, attirant souvent des associations entre variables. On trouve dans Santos (1981), Kalton et Kasprzyk (1982) et Little (1986a) des discussions des effets des imputations sur les relations.

La nature générale de l'effet de l'imputation sur les relations est mise en évidence si l'on examine son effet sur l'estimation de la covariance de l'échantillon dans la situation simple où la variable y a des réponses manquantes qui sont manquantes au hasard pour l'ensemble de la population et où la variable x n'a pas de données manquantes. La covariance de l'échantillon s_{xy} est calculée comme d'habitude, à partir des valeurs réelles pour les répondants et des valeurs imputées pour les non-répondants, comme une estimation de la covariance de la population S_{xy} . Si l'on utilise le fait que $E_2(y_{mis}) = y_{mid}$ comme ci-dessus, on peut montrer facilement que la valeur attendue de s_{xy} dans une méthode d'imputation déterministe est la même que celle en vertu de la méthode stochastique correspondante.

Comme le montre Santos (1980), le biais relatif de s_{xy} lorsqu'on utilise la méthode de la moyenne globale ou celle de l'imputation globale aléatoire est approximativement $-M$, où M est le taux de non-réponse. Cette situation s'explique par le fait que les valeurs y imputées ne sont pas reliées à leurs valeurs x , et que les cas avec des valeurs imputées atténuent la covariance vers zéro. Cette atténuation se trouve réduite en importance par les méthodes d'imputation qui utilisent les variables auxiliaires. Avec l'imputation de la moyenne de classe ou de l'imputation aléatoire intra-classe, le biais relatif est d'environ $-M(S_{xy,z}/S_{xy})$,

Cependant, cette façon d'agir comporte l'utilisation du modèle uniquement. Une autre solution qui éviterait l'hypothèse de normalité consisterait à choisir les résidus au hasard à partir de la distribution empirique des résidus des répondants. Une autre solution serait de choisir un résidu à partir d'un répondant qui correspond "de près" au non-répondant, avec une mesure "proche" en termes de valeurs semblables des variables auxiliaires. Cette solution intéressante évite l'hypothèse d'homoscédasticité et empêche une mauvaise spécification de distribution du terme résiduel. À la limite, le répondant le plus proche est celui qui a les mêmes valeurs de toutes les variables auxiliaires que le non-répondant. Dans ce cas, le non-répondant reçoit une des valeurs des répondants appartés. C'est ce qui se produit avec les méthodes hot-deck, lorsque les non-répondants et les répondants sont appartés en termes des variables auxiliaires et que les non-répondants se voient attribuer des valeurs provenant des répondants appartés.

Une autre considération dans le choix des résidus est de rendre les valeurs imputées plausibles. Comme on l'a noté plus haut, les méthodes déterministes peuvent imputer des valeurs pour les variables catégoriques et discrètes qui ne sont pas plausibles. Certaines méthodes stochastiques résolvent ce problème, par la répartition des résidus. En particulier, l'utilisation des résidus des répondants avec les méthodes aléatoire intra-classes et les méthodes hot-deck itérative et hiérarchique font en sorte que les valeurs imputées sont plausibles.

3.2 Classe d'imputation ou imputation par régression

Comme on l'a noté plus haut, les deux méthodes de classes d'imputation ou d'imputation par régression relèvent du modèle d'imputation donné par l'équation (2). La différence entre ces deux méthodes réside dans les façons dont elles emploient des variables auxiliaires.

Les méthodes de classes d'imputation divisent l'échantillon en un ensemble de classe. À cette fin, il faut catégoriser les variables auxiliaires continues. La constitution des classes est parfaitement flexible, et l'utilisation symétrique des variables auxiliaires dans différentes parties de l'échantillon n'est pas nécessaire. Ainsi, par exemple, lorsqu'on impute le taux de rémunération horaire dans un échantillon d'employés, on peut d'abord commencer par diviser l'échantillon en deux parties, soit les syndiqués et les non-syndiqués, ensuite on peut constituer les classes d'imputation pour les membres en termes d'âge et de profession, tandis que les classes non-syndiqués peuvent être constituées en termes de sexe et de branche d'activité. En règle générale, on vise à construire des classes de taille appropriée qui expliquent le plus possible la variance de la variable à imputer. Lorsque les classes sont constituées par une classification recoupée intégrale des variables auxiliaires, le modèle qu'on utilise, effectivement, contient tous les effets principaux et toutes les interactions pour la classification recoupée. Les méthodes de classe d'imputation sont limitées par le fait que le nombre de classes constituées doit être calculé de façon à garantir qu'il y a un nombre minimum de répondants dans chaque classe. La méthode hot-deck hiérarchique tente d'accroître la quantité de données auxiliaires utilisées, mais même avec cette méthode, il est souvent impossible de procéder à l'appariement des répondants et des non-répondants à des niveaux de détail plus fins. Combinées à l'utilisation des résidus des répondants aléatoires à l'intérieur d'une classe, les méthodes de classe d'imputation ont la propriété précieuse selon laquelle les valeurs imputées sont des valeurs acceptables, c'est-à-dire que les valeurs imputées sont les valeurs des répondants.

Les méthodes d'imputation par régression présentent un avantage par rapport à celles de classe d'imputation pour ce qui est du nombre et du niveau de détail des variables auxiliaires qu'elles peuvent utiliser. On peut prendre l'âge, par exemple, comme variable continue au lieu d'être catégorisée en quelques classes seulement. Le modèle de régression permet d'inclure davantage d'effets principaux dans le modèle, mais au prix d'une baisse du nombre d'interactions. Naturellement, les modèles de régression peuvent également inclure des termes polynomiaux et utiliser des transformations, mais là encore, il faut les préciser. Le modèle de régression présente le potentiel de donner de meilleures prédictions pour les valeurs imputées, mais pour cela il faut procéder à une modélisation minutieuse. Une modélisation minutieuse n'est pas réaliste pour toutes les variables d'une enquête, mais elle peut être réalisable pour une ou deux variables importantes, et en particulier dans le cas d'enquêtes périodiques. Sans

dans le cas de variables catégoriques ou discrètes (telles que être un membre de la population active (1) ou non (0), et le nombre d'années d'études). Les méthodes de la moyenne globale, de la moyenne de classe et d'imputation par régression imputent des valeurs telles que 0,7 pour un actif (c'est-à-dire une probabilité de 70%) et de 10,7 pour le nombre d'années d'études achevées. Ces valeurs ne conviennent pas pour des répondants individuels, et leur arrondissement à des nombres entiers se traduit par un biais. Pour cette raison, ces méthodes d'imputation ne donnent pas de bons résultats pour les variables catégoriques et discrètes. Un avantage appréciable de toutes les méthodes hot-deck est qu'elles donnent toujours des valeurs plausibles, puisque les valeurs proviennent des répondants.

Les méthodes d'imputation ci-dessus ont deux caractéristiques importantes sur lesquelles il faut s'arrêter un peu: lorsque l'on ajoute un résidu, et dans ce cas, sa forme, et si les données auxiliaires sont utilisées sous forme de variables dichotomiques pour représenter des classes, ou si elles sont utilisées directement dans la régression. Ces deux caractéristiques sont examinées dans les deux sous-sections suivantes. Les sous-sections suivantes examinent d'autres problèmes soulevés par l'utilisation de l'imputation.

3.1 Choix des résidus

On peut classer les méthodes d'imputation en déterministes ou stochastiques selon que les ϵ_{mi} sont posés comme étant égaux à zéro ou non. Pour chaque méthode d'imputation déterministes, il y a une contrepartie stochastique. Soit y^{mid} la valeur imputée par la méthode déterministe, et $y^{mis} = y^{mid} + \epsilon_{mi}$ celle imputée par la méthode stochastique correspondante. On obtient alors $E_2(y^{mis}) = y^{mid}$, où E_2 indique l'espérance mathématique pour l'échantillonage des résidus compte tenu de l'échantillon initial, à condition que $E_2(\epsilon_{mi}) = 0$ (ce qui est en général le cas).

Le choix entre une méthode déterministe et une méthode stochastique correspondante dépend de la nature de l'analyse d'enquête effectuée. Examinons d'abord l'estimation de la moyenne de la population de la variable y en utilisant pour cela la moyenne d'échantillon des valeurs des répondants et des valeurs imputées des non-répondants. Comme Kalton et Kasprzyk (1982) le montrent, avec $E_2(y^{mis}) = y^{mid}$, il s'en suit que l'espérance mathématique de la moyenne de l'échantillon est la même, quelle que soit la méthode d'imputation utilisée (déterministe ou stochastique). Les méthodes ont donc par conséquent le même effet sur le biais de l'estimation. Cependant, l'addition de résidus aléatoires dans la méthode stochastique entraîne une perte de précision dans la moyenne de l'échantillon. Bien qu'il soit possible de maîtriser cette perte en choisissant une méthode appropriée d'échantillonnage des résidus (Kalton et Kish 1984), il n'y en reste pas moins qu'il y a une certaine perte de précision. Pour cette raison, un schéma déterministe est préférable dans le cas de l'estimation de la moyenne de la population.

Considérons maintenant l'estimation de l'écart type des éléments et la distribution de la variable y . Les méthodes d'imputation déterministes ne donnent pas de bons résultats, puisque qu'elles entraînent une réduction de l'écart type et une déformation de la forme de la distribution. On peut illustrer simplement ceci en termes de la méthode d'imputation de la moyenne de classe. En attribuant la moyenne de classe à toutes les valeurs manquantes dans une classe, la forme de la distribution est indiscutablement déformée, avec une série de pointes comme moyennes de classe. L'écart type de la distribution se trouve atténué parce que les valeurs imputées ne prennent en compte que la variance inter-classes et non pas la variance intra-classes. L'intérêt des méthodes d'imputation stochastiques est que le terme résiduel saisit la variance inter-classes (ou résiduelle), et évite ainsi la réduction de l'écart type de l'élément et la distorsion de la distribution.

Comme certaines analyses d'enquête vont probablement faire intervenir les distributions des variables, on préfère en général des méthodes d'imputation stochastiques telles que les méthodes hot-deck. Une fois que l'on a décidé d'utiliser une méthode stochastique, la question du choix des résidus se pose. Si l'on accepte les hypothèses de régression habituelles, il est possible de choisir les résidus à partir d'une distribution normale avec une moyenne de zéro et une variance égale à la variance des résidus provenant de la régression des répondants.

réponse manquante est suivi d'un ou de plusieurs enregistrements avec des réponses manquantes. Le nombre de classes d'imputation qui peut être utilisé avec cette méthode doit également être limité afin de garantir que des donneurs existent à l'intérieur de chaque classe. On trouve des discussions utiles de la méthode hot-deck itérative dans Bailar *et al* (1978), Bailar et Bailar (1978, 1983), Ford (1983), Oh et Scheuren (1980), Oh *et al.* (1980), et Sande (1983).

(g) *Imputation hot-deck hiérarchique.* Cette méthode n'a pas les inconvénients mentionnés plus haut de l'imputation hot-deck itérative. Il s'agit d'une forme d'imputation hot-deck mise au point pour les questions de l'enquête supplémentaire sur le revenu de mars de la Current Population Survey. Cette méthode consiste à répartir les répondants et les non-répondants dans un grand nombre de classes d'imputation à partir d'une catégorisation détaillée d'un ensemble important de variables auxiliaires. On procède ensuite à l'appariement des non-répondants avec les répondants sur une base hiérarchique, en ce sens que si un appariement ne peut être fait dans la classe d'imputation initiale, les classes sont fusionnées et on fait l'appariement à un niveau de détail moins élevé. Coder (1978) et Weinik et Coder (1980) donnent d'autres renseignements sur cette procédure.

(h) *Imputation par régression.* Cette méthode utilise des données de répondants pour régesser la variable pour laquelle des imputations sont nécessaires sur un ensemble de variables auxiliaires. L'équation de régression sert ensuite à prédire les valeurs des réponses manquantes. La valeur imputée peut être soit la valeur prédite, soit la valeur prédite plus un résidu. Il y a plusieurs façons d'obtenir ce dernier, comme on le verra plus loin.

(i) *Appariement par fonction de distance.* Cette méthode attribue à un non-répondant la valeur du répondant "le plus proche", "le plus proche" étant défini en termes d'une fonction de distance pour les variables auxiliaires. Diverses formes de fonctions de distance ont été proposées (Sande 1979; Vacek et Ashikaga 1980; etc.), et la fonction peut être construite de façon à réduire l'utilisation multiple d'enregistrements donneurs en incorporant une pénalité pour chaque utilisation (Colledge *et al.* 1978).

Bien qu'à première vue ces procédés peuvent sembler assez disparates, ils peuvent être presque tous réunis à l'intérieur d'un cadre unificateur unique. Les méthodes peuvent toutes être décrites, au moins de façon approximative, comme des cas particuliers du modèle général de régression

$$y_{mi} = b_{ro} + \sum b_{rj} z_{mj} + e_{mi} \tag{2}$$

où y_{mi} est la valeur imputée de l'enregistrement i avec une valeur manquante y , z_{mj} sont les valeurs qui reflètent les variables auxiliaires pour l'enregistrement, b_{ro} et b_{rj} sont les coefficients de régression pour la régression de y sur x pour les répondants, et e_{mi} est un résidu choisi selon un schéma précisé pour la méthode d'imputation utilisée.

L'équation (2) représente la méthode d'imputation par régression de façon claire. Si les e_{mi} sont posés comme étant égaux à zéro, alors la valeur imputée est la valeur prédite à partir de la régression; sinon, on peut ajouter un résidu d'une forme quelconque. L'équation représente également l'imputation moyenne de classe en définissant les z_j comme des variables dichotomiques qui représentent des classes, et en posant $e_{mi} = 0$. L'équation de régression se réduit alors à $y_{mi} = \bar{y}_{rh}$, la moyenne de classe. L'imputation aléatoire intra-classe se fait en ajoutant un résidu à la moyenne de classe, résidu qui est l'écart par rapport à la moyenne de classe d'un des répondants. Dans ce cas, $y_{mi} = \bar{y}_{rh} + e_{rhk}$, où e_{rhk} est l'écart par rapport au répondant k dans la classe h ; ceci se ramène à $y_{mi} = \bar{y}_{rhk}$, qui est la valeur à la méthode aléatoire en question. Les méthodes hot-deck itérative et hiérarchique ressemblent à la méthode aléatoire intra-classe. La méthode de la moyenne globale et la méthode d'imputation globale aléatoire sont des cas dégénérés de la méthode de la moyenne de classe et de la méthode aléatoire intra-classe qui n'utilisent pas de données auxiliaires.

Une considération importante dans le choix des méthodes d'imputation est le type de variables que l'on impute. Toutes les méthodes ci-dessus peuvent être appliquées de façon courante en présence de variables continues, mais certaines d'entre elles ne conviennent pas

la base de cette façon d'agir est que les non-répondants sont comme les répondants tardifs. Cette hypothèse semble douteuse, cependant, et les preuves empiriques provenant d'une étude de suivi intensive des non-répondants dans l'U.S. Current Population Survey ne la confirment pas (Palmer et Jones 1966; Palmer 1967).

3. IMPUTATION

Un vaste éventail de méthodes d'imputation ont été mises au point pour attribuer les valeurs pour les réponses aux questions manquantes. On se contentera ici de présenter une brève vue d'ensemble des méthodes, les différences fondamentales entre elles et quelques-unes des questions soulevées par l'imputation. Kalton et Kasprzyk (1982) donnent un traitement plus poussé.

Les méthodes d'imputation peuvent aller de simples procédures *ad hoc* utilisées pour obtenir des enregistrements complets pour l'introduction des données aux techniques perfectionnées hot-deck et de régression. Voici quelques procédures d'imputation habituelles:

- (a) *Imputation déductive*. Il est parfois possible de déduire la réponse manquante à une question avec certitude à partir des réponses aux autres questions. Des vérifications doivent examiner la cohérence entre les réponses à des questions connexes. Lorsque les vérifications limitent une réponse manquante à une valeur possible seulement, on peut utiliser l'imputation déductive. L'imputation déductive est la forme idéale de l'imputation.
- (b) *Imputation par la moyenne globale*. Cette méthode attribue la moyenne globale des répondants à toutes les réponses manquantes.
- (c) *Imputation par la moyenne de classe*. L'échantillon total est divisé en classes selon les valeurs des variables auxiliaires servant à l'imputation (comparables aux classes de pondération). À l'intérieur de chaque classe d'imputation, la moyenne de classe des répondants est attribuée à toutes les réponses manquantes.
- (d) *Imputation globale aléatoire*. On choisit au hasard un répondant dans l'échantillon de répondants total, et la valeur du répondant choisi est attribuée au non-répondant. Cette méthode est la forme la plus simple de l'imputation hot-deck, c'est-à-dire une procédure d'imputation dans laquelle la valeur attribuée à une réponse manquante est prise chez un répondant à l'enquête courante.
- (e) *Imputation aléatoire intra-classe*. Dans cette méthode hot-deck, on choisit un répondant au hasard dans une classe d'imputation, et la valeur ainsi obtenue est attribuée au non-répondant.
- (f) *Imputation hot-deck itérative*. Ce terme sert ici à décrire la procédure utilisée pour les questions sur la population active de l'U.S. Current Population Survey (Brooks et Biallar 1978). La procédure commence par un ensemble de classes d'imputation. Une valeur unique pour la question faisant l'objet de l'imputation est attribuée à chaque classe (peut-être tirée d'une enquête antérieure). Les enregistrements du fichier de données de l'enquête sont alors examinés à leur tour. Si un enregistrement a une réponse à la question, la réponse obtenue remplace la valeur réservée pour la classe d'imputation à laquelle il appartient. Si l'enregistrement a une réponse manquante, il se voit attribuer la valeur réservée pour sa classe d'imputation.

La méthode hot-deck est semblable à l'imputation aléatoire intra-classe. Si l'ordre des enregistrements dans le fichier de données était aléatoire, les deux méthodes seraient équivalentes, à l'exception du début. L'ordre non aléatoire de la liste favorise en général la méthode hot-deck, puisqu'elle donne un appartement plus étroit des donneurs et des receveurs, à condition que l'ordre du fichier donne une autocorrélation positive. Les avantages sont, cependant, probablement peu appréciables.

La méthode hot-deck itérative présente l'inconvénient de pouvoir facilement trouver des utilisations multiples pour les donneurs, caractéristique qui se traduit par une perte de précision dans les estimations d'enquête. Une telle utilisation multiple des donneurs se produit lorsque, à l'intérieur d'une classe d'imputation, un enregistrement avec une

sous-ensembles aléatoires des populations de la cellule. Le deuxième terme est zéro si $W_{hk} = W_{hk}$, ou s'il n'y a aucune interaction dans Y_{rhk} pour la classification. À la base de la méthode du quotient se trouve un modèle logit pour les taux de réponse de cellules. Avec le modèle $\ln[R_{hk}/(1 - R_{hk})] = \alpha_h + \beta_k$ pour les taux de réponse dans une classification bidimensionnelle, $W_{hk} = W_{hk}$. Par conséquent, dans ce modèle, le deuxième terme dans $B(Y^q)$ est zéro.

La pondération par la méthode du quotient est examinée plus en détail par Oh et Scheuren (1978a, 1978b, 1983). Oh et Scheuren (1987a) donnent également une bibliographie sur la méthode itérative du quotient.

2.4 Pondération avec probabilités de réponse

Bien qu'un certain nombre de méthodes pour la pondération avec des probabilités de réponse aient été proposées, cette approche n'a pas encore été adoptée de façon générale comme procédure de correction. À la base de cette méthode, on suppose que tous les éléments de la population ont des probabilités (habituellement devant être différentes de zéro) de répondre à l'enquête. On utilise une méthode pour estimer les probabilités de réponse des éléments répondants. Ces derniers se voient à leur tour attribuer des poids de correction de la non-réponse inversement proportionnels à leurs probabilités de réponse estimatives.

Une des premières applications de cette méthode est le procédé bien connu de Politz et Simmons (1949, 1950). Un contact simple (le soir) est pris pour chaque ménage sélectionné, et au cours de l'interview, on demande aux répondants pendant combien des cinq soirs précédents ils se trouvaient à la maison à peu près à la même heure. Leurs probabilités de réponse sont ensuite calculées comme étant une fraction des six soirs (y compris celui de l'interview) pendant lesquels ils étaient à la maison, et les inverses de ces probabilités servent à l'analyse. Il est à noter que cette méthode ne prend pas en compte ceux qui étaient absents les six soirs, et ceux qui ont refusé de répondre.

Une autre façon d'estimer les probabilités de réponse est de procéder à la régression du statut de réponse (1 pour les répondants, 0 pour les non-répondants) sur un ensemble de variables existant pour les répondants et les non-répondants, en utilisant pour cela une régression logistique ou probit. Les valeurs prédites à partir de la régression pour les répondants sont ensuite retenues comme étant leurs probabilités de réponse, et les poids inversement proportionnels à ces valeurs prédites servent à l'analyse. Un cas spécial se pose lorsque les variables explicatives sont des variables dichotomiques qui identifient un ensemble de classes. Les probabilités de réponse prédites sont alors les taux de réponse de classe, et la méthode se ramène à une correction des pondérations de l'échantillon. La méthode convient le mieux aux situations lorsque l'on dispose de beaucoup d'informations pour les non-répondants, comme par exemple lorsque les non-répondants sont des pertes après la première vague d'une enquête par panel. Little et David (1983) ont examiné la possibilité d'appliquer cette méthode à la non-réponse par panel. Il convient de noter que si la régression prédit très bien le statut de réponse, les poids en résultant pourront varier de façon appréciable, ce qui se traduira par une perte sensible de la précision des estimations de l'enquête.

Drew et Fuller (1980, 1981) présentent une approche pour l'estimation des probabilités de réponse à partir du nombre de répondants certains à des contacts successifs. Dans leur modèle, la population est divisée en classes. À l'intérieur de chacune d'elle, chaque élément par hypothèse reçoit la même probabilité de réponse, qui demeure la même pour chaque contact. Le modèle prévoit également une proportion de non-répondants enduits que l'on suppose constante pour les différentes classes. Dans ces hypothèses, les probabilités de réponse de chaque classe et la proportion de non-répondants enduits peuvent être estimées, et par conséquent il est possible d'apporter des corrections des pondérations. Thomsen et Siring (1983) ont adopté une méthode semblable, utilisant un modèle plus complexe.

Enfin, il faut mentionner une approche voisine qui compense la non-réponse par une hausse de la pondération des répondants difficiles à interviewer. Bartholomew (1961), par exemple, a proposé de ne faire que deux contacts dans une enquête, et de pondérer à la hausse les répondants lors du deuxième appel afin de représenter les non-répondants. L'hypothèse à

situation, on limite habituellement le degré de segmentation de l'échantillon. Mais dans ce dernier cas, il peut y avoir encore des classes de pondération nécessitant des poids élevés. Parfois, ce problème est traité par fusionnement avec des classes voisines, et parfois leurs poids sont réduits à une valeur maximum acceptable quelconque (voir Bailar *et al.* 1978, et Chapman *et al.* 1986, pour des exemples). Ces procédés évitent l'augmentation de la variance associée à l'utilisation de poids extrêmes, mais ils peuvent se traduire par une augmentation du biais; leur effet sur ce dernier est cependant inconnu.

Dans certains cas, il semble souhaitable d'utiliser plusieurs variables auxiliaires dans la constitution des classes de pondération pour les corrections des pondérations de la population ou de l'échantillon. Cependant, si les classes sont constituées en prenant la classification recoupée intégrale des variables, on se trouvera en présence d'un grand nombre de classes de pondération. À moins que l'échantillon ne soit très grand, les tailles des échantillons dans les classes de pondération obtenues seront petites, et l'instabilité des taux de réponse se traduira par une importante variance des poids et par une perte de précision des estimations d'enquête. Une façon de traiter ce problème est de réduire le nombre de classes en fusionnant des cellules, en laissant de côté par exemple quelques-unes des variables auxiliaires, ou en-core, en utilisant des classifications plus grossières. Une autre solution consisterait à baser les poids sur un modèle, comme on le fait dans le cas de la méthode itérative du quotient, examinée ci-dessous.

2.3 Corrections selon la méthode itérative du quotient

Lorsque les classes de pondération sont choisies comme étant des cellules de la classification recoupée des variables auxiliaires, les corrections des pondérations de la population rendent la distribution conjointe des variables auxiliaires de l'échantillon conforme à celle de la population. De même, les corrections des pondérations de l'échantillon rendent la distribution conjointe des variables auxiliaires de l'échantillon de répondants conforme à celle de l'échantillon total. Comme on l'a dit plus haut, cependant, cette méthode peut avoir l'effet indésirable de créer un grand nombre de petites classes de pondération, qui sont par conséquent instables. Par ailleurs, il n'est pas toujours possible d'utiliser cette méthode dans les corrections des pondérations de la population; souvent, les distributions marginales de la population, et peut-être, quelques distributions bivariées, des variables auxiliaires existent, mais la distribution conjointe intégrale est inconnue.

Une autre solution serait de définir des poids qui rendent les distributions marginales des variables auxiliaires de l'échantillon conformes aux distributions marginales de la population (pondérations de la population) ou aux distributions d'échantillon totales marginales (pondérations de l'échantillon), sans garantir pour cela la conformité de la distribution conjointe intégrale. La méthode itérative du quotient, ou plus simplement, méthode du quotient, peut être utilisée pour obtenir des poids qui répondent à ces conditions. La méthode du quotient correspond à l'ajustement proportionnel itératif dans l'analyse des tableaux de contingence (voir, par exemple, Bishop *et al.* 1975).

Considérons l'utilisation de la méthode du quotient dans le cas simple de deux variables auxiliaires. Soit W_{hk} la proportion de la population dans la cellule (h, k) de la classification recoupée, et soit w_{hk} la proportion attribuée à cette cellule par l'algorithme de la méthode du quotient. Suivant les tailles de l'échantillon total et de celui des répondants dans les cellules, et en supposant que toutes les cellules ont au moins un répondant, le biais de la moyenne de l'échantillon corrigé de la méthode du quotient $\bar{y}^q = \sum w_{hk} \bar{y}_{hk}$ est

$$B(\bar{y}^q) = \sum W_{hk} M_{hk} (\bar{y}_{rhk} - \bar{y}_{mhk}) + \sum (W_{hk} - w_{hk}) (\bar{y}_{rhk} - \bar{y}_{rh.} - \bar{y}_{rk.} + \bar{y}_{r.})$$

où $W_{hk} = E(w_{hk})$. Le premier terme de ce biais correspond au terme du biais B dans l'équation (1) pour les corrections des pondérations de la population et de l'échantillon. L'espérance mathématique de ce terme est zéro si les non-répondants de la cellule sont des

variables (urbaine/rurale, région géographique) et parfois, à quelques variables supplémentaires se trouvant dans la base de sondage. À l'occasion, il peut également être possible de recueillir des informations sur une ou deux variables pour les non-répondants, comme par exemple par observation de l'interviever.

Tout comme les corrections des pondérations de la population ressemblent à la post-stratification, de même les corrections des pondérations de l'échantillon ressemblent à l'échantillonnage à deux degrés. L'échantillon du premier degré est l'échantillon total des répondants et des non-répondants, l'échantillon du deuxième degré est le sous-échantillon des répondants, sélectionné avec différentes fractions de sondage (taux de réponses) dans différentes strates (classes de pondération). La moyenne pondérée de l'échantillon peut être représentée par la formule $\bar{y}_s = \sum w_h y_{rh}$, où w_h est la proportion de l'échantillon total dans la classe de pondération h . En supposant qu'il n'y a pas d'erreurs d'observation, $E(w_h) = W_h$, la proportion de la population dans la classe h , telle qu'elle est utilisée dans l'estimateur pondéré $B(\bar{y}_s) = B$ d'après l'équation (1). Par conséquent, l'effet de la correction des pondérations de l'échantillon sur le biais de l'estimation d'enquête est le même que celui de la correction des pondérations de la population. Comme les corrections de pondérations de l'échantillon n'utilisent que des données pour l'échantillon, elles ne corrigent pas les erreurs d'observation (contrairement aux corrections des pondérations de la population).

Les deux types de correction ont des besoins en données différents, et reçoivent ainsi des causes possibles différentes de biais. Dans la pratique, les deux types de correction sont utilisés en combinaison. En règle générale, les corrections des pondérations de l'échantillon sont appliquées d'abord, et les corrections des pondérations de la population, ensuite. Une façon habituelle de procéder est d'abord de déterminer les poids d'échantillon nécessaires pour corriger les probabilités inégales de sélection, ensuite de réviser ces poids afin de corriger les taux de réponse inégaux dans différentes classes de pondération d'échantillon (classes urbaines/rurales à l'intérieur de régions géographiques, par exemple), et enfin, de réviser les poids une nouvelle fois afin d'aligner la distribution d'échantillon pondérée pour certaines caractéristiques. L'utilisation de cette méthode dans la U.S. Current Population Survey est décrite par Bailar *et al.* (1978).

Tout comme pour les corrections des pondérations de la population, les corrections des pondérations de l'échantillon visent à réduire le biais que la non-réponse pourrait causer dans les estimations d'enquête. Un des effets des corrections des pondérations de l'échantillon est d'augmenter les variances des estimations d'enquête. Il faut par conséquent trouver un compromis entre la réduction du biais et l'augmentation de la variance. On peut avoir une idée de l'accroissement de la variance imputable à la pondération en examinant le cas où les variances des éléments à l'intérieur des classes de pondération sont toutes les mêmes et les variances entre les moyennes de classe sont négligeables lorsqu'on les compare aux variances intra-classes. Alors, la perte de précision résultant de la pondération est à peu près la même que celle provenant de l'utilisation d'un échantillon stratifié disproportionné, alors qu'un échantillon stratifié proportionné est optimal; Kish (1965, Section 11.7C; 1976) examine ce dernier cas.

Dans ces conditions, la pondération accroît la variance de la moyenne d'un échantillon d'approximativement $L = (\sum W_h^h k_h) (\sum W_h / k_h)$, où W_h est la proportion de la population, et k_h est le poids de la classe h . Une autre forme de L est $(\sum n_h) (\sum n_h k_h^2) / (\sum n_h k_h)^2$, où n_h est la taille de l'échantillon dans la classe h . Le facteur L devient grand lorsque la variance des poids est importante.

Une variance importante dans les poids peut résulter de la segmentation de l'échantillon en un grand nombre de classes de pondération avec peu d'éléments échantillonnés dans chacune. Lorsque les classes de pondération sont peu nombreuses, leurs taux de réponse sont instables, ce qui se traduit par une forte variation des poids. Afin d'éviter cette

probabilité égale. Supposons que la population est divisée en un ensemble de classes de pondération, avec une proportion W_h d'éléments dans la classe h . Supposons de plus que les répondants répondent à chaque fois, et que les non-répondants n'existent pas. Soit R_h et M_h les proportions des répondants et non-répondants respectivement de la classe h , et soit $\bar{R} = \sum W_h R_h$ le taux de réponse global. Alors, d'après Thomsen (1973), le biais de la moyenne des répondants non corrigée (\bar{y}) peut s'écrire sous la forme

$$(1) \quad B(\bar{y}) = \bar{R}^{-1} \sum W_h (Y_{rh} - Y_r) (R_h - \bar{R}) + \sum W_h M_h (Y_{rh} - Y^{mh}) = A + B$$

où Y_{rh} et Y^{mh} sont les moyennes des répondants et des non-répondants de la classe h respectivement, et Y_r est la moyenne de la population des répondants. L'emploi de la correction des pondérations de la population donne la moyenne de l'échantillon pondérée, $\bar{y}_p = \sum W_h Y_{rh}$, où \bar{y}_{rh} est la moyenne de l'échantillon de répondants de la classe h . Le biais de \bar{y}_p est simplement le second terme de $B(\bar{y})$, c'est-à-dire $B(\bar{y}_p) = B$.

Si A et B sont de même signe, la correction des pondérations de la population réduit de A le biais absolu dans l'estimation de Y . Si $Y^{rh} = Y^{mh}$, ce qui se produit lorsqu'on s'attend à ce que les non-répondants soient absents au hasard à l'intérieur des classes de pondération, alors $B = 0$. Dans ce cas, la correction des pondérations de la population élimine le biais. Le terme A est un terme de genre covariancé entre les taux de réponses de classe et les moyennes des répondants de classe. Il sera égal à 0 si le taux de réponse ou la moyenne des répondants ne varie pas entre classes. Dans l'un ou l'autre de ces cas, la correction des pondérations de la population n'a aucun effet sur le biais de l'estimateur. On peut observer que les corrections des pondérations de la population peuvent accroître le biais absolu de l'estimation de Y . Ceci sera le cas lorsque A et B seront de signes opposés, et si $|A| < 2|B|$. Les corrections des pondérations de la population nécessitent des données extérieures sur les distributions de la population pour les variables à utiliser. Il faut prendre soin de s'assurer que les données sur lesquelles reposent les distributions de population soient exactement comparables aux données d'enquête; sinon, on obtiendra des poids qui ne conviennent pas. Comme la procédure touche les distributions des populations, elle fait plus que simplement de corriger la non-réponse. Elle compense les erreurs de champ d'observation et apporte une correction de post-stratification.

2.2 Correction des pondérations de l'échantillon

Comme pour les corrections des pondérations de la population, les corrections des pondérations de l'échantillon consistent à diviser l'échantillon en classes de pondération; des poids différents sont alors affectés à ces classes afin d'essayer de réduire le biais de non-réponse. La différence essentielle entre les deux procédures réside dans les informations auxiliaires utilisées. Comme l'a décrit plus haut, les corrections des pondérations de la population sont basées sur des distributions de population d'origine externe. Aucune donnée n'est nécessaire pour les non-répondants d'échantillon. Par contre, les corrections des pondérations de l'échantillon n'utilisent que des données internes à l'échantillon et nécessitent des informations sur les non-répondants.

Dans le cas des corrections des pondérations de l'échantillon, les poids des corrections de non-réponse des classes de pondération sont établis de façon proportionnelle aux inverses des taux de réponse des classes. Afin de calculer ces taux de réponses, il faut déterminer le nombre de répondants et de non-répondants dans les classes. Il faut par conséquent connaître à quelle classe chaque répondant et non-répondant appartient. Comme en règle générale on ne dispose que de très peu d'informations sur les non-répondants, le choix de la classe de pondération est habituellement très sérieusement limité. Il se trouve souvent limité à des variables de plans de sondage généraux (UPB et strates), aux caractéristiques de ces

répondant met fin prématurément à une interview, lorsque des données ne sont pas recueillies pour un ou plusieurs membres d'un ménage autrement coopératif (pour l'analyse au niveau des ménages) ou lorsqu'une personne échantillonnée fournit des données pour une partie seulement des questions d'un panel. Cox et Cohen (1985) et Kalton (1986) examinent le choix entre la pondération et l'imputation pour compenser la non-réponse lors d'une enquête par panel. Bien que les corrections de pondérations et d'imputation soient traitées comme des approches distinctes dans la discussion qui suit, elles sont en fait étroitement reliées. La relation et les différences entre ces deux approches sont examinées brièvement à la section 4, qui contient également quelques autres solutions au problème des données d'enquête manquantes.

2. CORRECTIONS DES PONDERATIONS

Ces corrections sont destinées avant tout à compenser la non-réponse intégrale. L'essence de toute procédure de correction de pondérations est d'accroître les poids de répondants précis de façon à ce que ces derniers représentent des non-répondants. Les procédures nécessitent des renseignements auxiliaires sur les non-répondants ou la population totale. Les quatre types suivants de correction des pondérations sont examinés brièvement ci-dessous: corrections des pondérations de la population, correction des pondérations de l'échantillon, méthode itérative du quotient et poids basés sur les probabilités de réponse. Kalton (1983) contient davantage de détails.

2.1 Corrections des pondérations de la population

L'information auxiliaire utilisée dans la correction des pondérations de la population est la distribution de la population sur une ou plusieurs variables telles que la répartition de la population selon l'âge, le sexe et la race obtenues à partir d'estimations démographiques types. L'échantillon des répondants est partagé en un ensemble de classes, appelées ici classes de pondération, définies par l'information auxiliaire disponible (hommes blancs âgés de 15-24 ans, femmes non blanches âgées de 25-34 ans, etc.). Les poids de tous les répondants à l'intérieur d'une classe de pondération sont ensuite corrigés par le même facteur multiplicatif avec différents facteurs dans différentes classes. La correction est faite de façon à ce que la distribution des répondants pondérée pour l'ensemble des classes de pondération soit conforme à la distribution de la population.

On appelle souvent ce type de correction post-stratification. Nous évitons ce terme ici, cependant, parce que même si la pondération de population ressemble à la post-stratification, il existe une importante différence entre les deux. Tout comme la pondération de population, la post-stratification pondère l'échantillon de façon à rendre la distribution d'échantillonnage conforme à la distribution de la population pour un ensemble de classes (ou strates). Cependant, la théorie habituelle de la post-stratification ne porte que sur les fluctuations d'échantillonnage qui pourraient pousser la distribution d'échantillonnage à s'écarter de la distribution de la population, et non pas sur les écarts plus importants qui peuvent résulter de taux de réponse différents pour l'ensemble des classes. Les corrections de post-stratification ressemblent davantage à un affinement de l'échantillon, se traduisant en général par de faibles variations seulement des poids à travers les strates. Par conséquent, et à condition que les strates ne soient pas petites, la post-stratification se traduit par des erreurs types moins élevées pour les estimations d'enquête. Par contre, les corrections des pondérations de la population peuvent entraîner davantage de corrections majeures et causer des erreurs types plus élevées.

Les corrections des pondérations de la population tentent de réduire le biais entraîné par la non-réponse et les erreurs d'observation. Considérons l'estimation d'une moyenne de la population Y à partir d'un échantillon dans lequel les éléments sont sélectionnés avec une

Le traitement des données d'enquête manquantes

GRAHAM KALTON et DANIEL KASPRZYK¹

RÉSUMÉ

La non-réponse intégrale et la non-réponse à une question sont les causes des données d'enquête manquantes. On corrige habituellement la non-réponse intégrale par une certaine correction des pondérations, tandis que la non-réponse à une question est corrigée par une sorte d'imputation. Les auteurs examinent les méthodes de correction par pondération et imputation ainsi que leurs propriétés.

MOTS CLÉS: Non-réponse; non-réponse à une question; corrections de pondération; imputation.

1. INTRODUCTION

En règle générale, les enquêtes recueillent les réponses à un grand nombre de questions pour chaque élément échantillonné. Le problème des données manquantes se pose lorsque une partie ou la totalité des réponses ne sont pas recueillies pour un élément échantillonné, ou lorsque certaines réponses sont supprimées parce qu'elles ne répondent pas aux critères de vérification. On distingue habituellement entre la non-réponse intégrale (ou à une question), quand aucune réponse à l'enquête n'existe pour un élément échantillonné, et la non-réponse à une question, quand quelques-unes des réponses, mais pas toutes, sont données. La non-réponse intégrale résulte de refus, de l'incapacité à participer à l'enquête, d'absences et d'éléments non relevés. La non-réponse à une question s'explique par des refus, l'ignorance, des omissions et des réponses supprimées lors de la vérification.

On examine dans cette communication des méthodes générales qui existent pour le traitement des données d'enquête manquantes. La distinction entre non-réponse intégrale et non-réponse à une question est utile ici, puisqu'on utilise des méthodes de correction différentes dans les deux cas. En général, la seule information existante à propos de l'ensemble des non-répondants est celle concernant la base de sondage d'où l'échantillon a été tiré (les strates et les UPÉ dans lesquelles ils se trouvent, par exemple). Il est habituellement possible d'incorporer rapidement les aspects importants de ces renseignements dans des corrections de pondérations afin d'essayer de compenser les données manquantes. On se sert habituellement des corrections de pondérations en règle générale pour la non-réponse intégrale. On examine à la section 2 les méthodes pour faire de telles corrections.

Par contre, dans le cas de la non-réponse à une question, on dispose d'un grand nombre d'informations supplémentaires pour les éléments en cause, c'est-à-dire non seulement les renseignements provenant de la base de sondage, mais également les réponses aux autres questions de l'enquête. Afin de conserver toutes les réponses pour des éléments contenant des non-réponses à des questions, la méthode habituelle de correction donne des enregistrements d'analyse qui incorporent des réponses réelles aux questions pour lesquelles les réponses ont été acceptables, et des réponses imputées aux autres questions. Les méthodes d'imputation pour l'affectation des réponses aux questions manquantes sont examinées à la section 3. En général, le choix entre les corrections de pondérations et l'imputation pour le traitement des données d'enquête manquantes est assez aisé, mais le choix n'est pas toujours aussi clair. On peut citer des cas de ce que l'on pourrait appeler une non-réponse partielle, lorsque certaines données sont obtenues pour un élément échantillonné, mais une quantité appréciable de données est manquante. La non-réponse partielle peut se produire par exemple lorsque un

¹ Graham Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48106-1248 et Daniel Kasprzyk, Population Division, U.S. Bureau of the Census, Washington, D.C., 20233. Les auteurs aimeraient remercier les critiques pour leurs commentaires utiles.

PREFACE

Ce numéro renferme des articles présentés au Symposium de méthodologie sur les données manquantes dans les enquêtes, qui a eu lieu à Statistique Canada à Ottawa les 16 et 17 avril, 1986. Le symposium a été parrainé par le Comité d'étude des méthodes de Statistique Canada et le Laboratoire de recherche en statistique et en probabilité de la Carleton University. L'intérêt manifesté pour les données manquantes dans les enquêtes (en raison de la non-réponse ou des réponses inutilisables) s'est accru au cours des dernières années. Lors du symposium, plus de 200 professionnels des universités, des organismes gouvernementaux et des entreprises privées du Canada et des Etats-Unis ont eu l'occasion de s'informer des développements récents, théoriques ainsi que pratiques.

Le statisticien en chef du Canada, M. Ivan Fellegi a prononcé l'allocation d'ouverture. Il a fait état des inquiétudes soulevées à l'échelle mondiale à propos du fossé croissant qui sépare la statistique théorique de la statistique appliquée et a félicité les organisateurs d'avoir rassemblé des spécialistes de ces deux domaines. Tout en affirmant que le symposium visait principalement à approfondir la question des données manquantes, M. Fellegi a aussi reconnu que l'on pouvait profiter de l'occasion pour se demander dans quelle mesure les organismes statistiques devraient participer à l'établissement de modèles.

Le symposium comprenait quatre sessions. La première session, "Questions générales et expériences organisationnelles", a été présidée par L. Kish de l'University of Michigan et incluait des présentations de G. Kalton (University of Michigan), G.B. Gray (Statistique Canada), D.W. Chapman (U.S. Bureau of the Census) et L.R. Curtin (U.S. National Center for Health Statistics). Le président de la session de l'après-midi du 16 avril, "Le plan de sondage et l'estimation" était M. Hansen du Westat Inc. Des communications ont été présentées par P.S.R.S. Rao (University of Rochester), S. Michaud (Statistique Canada), C.E. Särndal (Université de Montréal), G. Lazarus (Statistique Canada) et V.P. Godambe (University of Waterloo).

La séance du matin du 17 avril, "La non-réponse et l'imputation des postes d'un questionnaire" a été présidée par M. Moore de l'Université de Montréal. Cette session comprenait les présentations de D. Rubin (Harvard University), P. Giles (Statistique Canada), M.S. Srivastava (University of Toronto) et M.A. Hidiroglou (Statistique Canada). Le président de la session finale, "Etudes de cas", était J.N.K. Rao de la Carleton University. Des communications de S. Hinkins (U.S. Internal Revenue Service), V. Tremblay (Université de Montréal) et S. Cheung (Statistique Canada) ont été présentées. Le symposium s'est terminé par une discussion générale des faits nouveaux concernant les données manquantes dans les enquêtes menée par J.N.K. Rao (président), G. Kalton, L. Kish, D. Rubin et I. Sande (Statistique Canada).

Ce numéro de la revue inclut neuf des articles du symposium. D'autres articles du symposium ayant été acceptés pour publication vont apparaître dans le prochain numéro.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 12, numéro 1, juin 1986

Edition spéciale – Les données manquantes dans les enquêtes

TABLe DES MATIÈRES

Preface	
G. KALTON et D. KASPRZYK	
Le traitement des données d'enquête manquantes	1
R. PLATEK et G.B. GRAY	
Sur les définitions des taux de réponse	19
V.P. GODAMBE et M.E. THOMPSON	
Résultats optimaux en situation de non-réponse	31
D.B. RUBIN	
Initiation à l'imputation multiple pour les cas de non-réponse	41
P. GILES et C. PATRICK	
Méthodes d'imputation dans un système généralisé	53
M.S. SRIVASTAVA et E.M. CARTER	
Application de la méthode du maximum de vraisemblance au traitement de la non-réponse dans les enquêtes par sondage	67
M.A. HIDIROGLOU et J.-M. BERTHELOT	
Contrôle statistique et imputation dans les enquêtes-entreprenises périodiques	79
V. TREMBLAY	
Critères pratiques pour la définition des classes de pondération	91
S. CHEUNG et C. SEKO	
Étude des effets des groupes d'imputation dans la méthode d'imputation du plus proche voisin pour l'enquête nationale sur les fermes	105

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics.

COMITÉ DE RÉDACTION

Président

R. Platek, *Statistique Canada*

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*

D.R. Bellhouse, *University of Western*

Ontario

L. Biggert, *Université de Florence*

E.B. Dagum, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

G.J.C. Hole, *Statistique Canada*

Rédacteurs adjoints

J. Armstrong, *Statistique Canada*

H. Lee, *Statistique Canada*

COMITÉ DE DIRECTION

R. Platek (Président), J. Armstrong, E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publiera des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception d'échantillon de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière sera accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociale, Statistique Canada, 4^e étage, Édifice Jean-Talon, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 10,00\$ par copie, 20,00\$ par année au Canada, et de 11,50\$ par copie, 23,00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. (Un prix réduit est offert aux membres de certaines organisations statistiques. Veuillez communiquer avec votre organisation pour vérifier si vous pouvez vous abonner aux prix réduits et envoyer votre demande d'abonnement directement à l'organisation.)

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA
JUIN 1986

Publication autorisée par
le ministre des Approvisionnements
et Services Canada

©Ministre des Approvisionnements
et Services Canada 1986

Novembre 1986
8-3200-501

Prix: Canada, \$10.00, \$20.00 par année
Autres pays, \$11.50, \$23.00 par année

Paiement en dollars canadiens ou l'équivalent
Catalogue 12-001, vol. 12, n° 1

ISSN 0714-0045

Ottawa

Canada

VOLUME 12, NUMÉRO 1
JUN 1986

UNE REVUE
DE
STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



12-001



Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 12, NUMBER 2
DECEMBER 1986

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

DECEMBER 1986

Published under the authority of
the Minister of Supply and
Services Canada

©Minister of Supply
and Services Canada 1987

March 1987
8-3200-501

Price: Canada, \$10.00, \$20.00 a year
Other Countries, \$11.50, \$23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 12, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

EDITORIAL BOARD

Chairman R. Platek, *Statistics Canada*
Editor M.P. Singh, *Statistics Canada*

Associate Editors

K.G. Basavarajappa, <i>Statistics Canada</i>	T.M. Jeays, <i>Statistics Canada</i>
D.R. Bellhouse, <i>University of Western Ontario</i>	G. Kalton, <i>University of Michigan</i>
L. Biggeri, <i>University of Florence</i>	C. Patrick, <i>Statistics Canada</i>
E.B. Dagum, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
W.A. Fuller, <i>Iowa State University</i>	C.E. Särndal, <i>University of Montreal</i>
J.F. Gentleman, <i>Statistics Canada</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
G.J.C. Hole, <i>Statistics Canada</i>	V. Tremblay, <i>Statplus, Montreal</i>
	K.M. Wolter, <i>U.S. Bureau of the Census</i>

Assistant Editors

J. Armstrong, *Statistics Canada* H. Lee, *Statistics Canada*

MANAGEMENT BOARD

R. Platek (Chairman), J. Armstrong, E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$10.00 per copy, \$20.00 per year in Canada, \$11.50 per copy, \$23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. (Reduced prices are available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.)

SURVEY METHODOLOGY

A Journal of Statistics Canada

Volume 12, Number 2, December 1986

CONTENTS

J.G. KOVAR

Estimating a Monthly Index Based on Trimestrial Data 107

A. TEN CATE

Regression Analysis Using Survey Data with Endogenous Design 121

D.A. BINDER and G. LAZARUS

A Cluster Analysis of Activities of Daily Living from the Canadian Health
and Disability Survey 139

G. HUOT and N. GAIT

Additive Versus Multiplicative Seasonal Adjustment When There Are Fast
Changes in the Trend-Cycle 151

Special Section – Missing Data in Surveys*

D.W. CHAPMAN, L. BAILEY, and D. KASPRZYK

Nonresponse Adjustment Procedures at the U.S. Bureau of the Census 161

S. HINKINS and F. SCHEUREN

Hot Deck Imputation Procedure Applied to a Double Sampling Design 181

S. MICHAUD

Comparison of Weighting and Imputation Methods for Estimating Unsampled
Data 197

C.E. SÄRNDAL

A Regression Approach to Estimation in the Presence of Nonresponse 207

P.S.R.S. RAO

Ratio Estimation with Subsampling the Nonrespondents 217

Acknowledgements 231

* The June 1986 issue was entirely devoted to selected papers presented at the Symposium on Missing Data in Surveys. Due to space limitations in the June issue, some symposium papers are included here.

Estimating a Monthly Index Based on Trimestrial Data

JOHN G. KOVAR¹

ABSTRACT

A problem of estimating monthly movements in rents based on data collected every four months is explored. Five alternative composite estimators of the rent index are presented and justified, both from an intuitive as well as theoretical point of view. An empirical study testing and comparing the proposed methods is described and summarized. Recommendations are put forth.

KEY WORDS: Index numbers; Rotating samples; Composite estimation.

1. INTRODUCTION

The rent component of the Consumer Price Index is based on data collected on a six month rotating basis using a Labour Force Survey Supplement. Since changes in rents generally occur on an annual basis, the effective sample size of the Labour Force Survey design is reduced. Furthermore, special annual benchmarks, which are obtained by revisiting the June sample of dwellings one year later, indicate that the rent component can suffer from varying degrees of bias (Dolson 1982). To ameliorate the situation, several data collecting schemes were proposed in order to combine the monthly data with the yearly benchmarks in a continuous and timely fashion. One of these methods, which collects data every four months, was selected for practical application.

The proposed design consists of four sets of four rotation groups of rented dwellings, each set of which is to be surveyed in one of four consecutive months, on a rotating basis. Each month, one rotation group is surveyed for the first time and the other three are those that rotated in four, eight and twelve months ago respectively. Each group would thus be surveyed four times over a period of thirteen months, before rotating out of the sample. Every month, data on current rents, as well as matched rents collected four months ago, are available from exactly three rotation groups (the fourth group is new and thus has no matching "backrents"). Yearly benchmarks can be calculated monthly based on one rotation group. This paper discusses several methods of estimating a monthly index based on such trimestrial data.

In estimating the indices, the constraints of the Consumer Price Index publication policy must be kept in mind. In other words, it must be practically as well as technically possible to produce the indices on a monthly basis for each of the index cities. The estimates must be timely: produced no later than mid-month following the reference month. Furthermore, no revisions can be made once the indices are published. While not entirely essential, it would be desirable that any proposed estimator be able to reflect (real) sudden changes in trend very quickly. On the other hand, in order to remain credible, the indices must be relatively stable: volatile, saw-toothed indices are to be avoided.

¹ John G. Kovar, Business Survey Methods Division, Statistics Canada, 11th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

In Section 2, five estimators will be presented, justified, and compared on a theoretical basis. Some empirical adjustments to these indices will be discussed in Section 3. In order to compare the performance of these estimators over time and between locations, a simulation study involving eight cities with observations over a period of 48 months was performed. The results of the study are presented in Section 4. The conclusions and recommendations can be found in Section 5.

2. INDEX ESTIMATORS

In this paper, only matched indices will be considered. While relative changes could easily be derived by comparing independent (unmatched) estimates of rent levels at distinct time points, such estimates of levels would have to be very reliable, necessitating prohibitively large sample sizes. Moreover, past studies indicate that such direct estimators tend to be volatile, upwardly biased and generally not practical in use (Szulc 1983). In what follows, therefore, an estimate of relative change between two time points will be based only on those units that report rents for both of these time points.

We will denote by x_m the total rent paid, in the current month m , by a certain subset s of dwellings in a given city. Thus, more rigorously,

$$x_m = \sum_{i \in s} x_{mi}, \quad (2.1)$$

where x_{mi} denotes the rent paid by the i -th dwelling in month m . The rent index is customarily estimated by chaining one month relatives, that is, the ratios of average rents between two consecutive months denoted by r_{m-1}^m . In other words, the index in month m , I_m , over a base period zero, is estimated recursively by

$$\hat{I}_m = \hat{I}_{m-1} \times \hat{r}_{m-1}^m = 100 \times \hat{r}_0^1 \times \hat{r}_1^2 \times \dots \times \hat{r}_{m-2}^{m-1} \times \hat{r}_{m-1}^m. \quad (2.2)$$

where 100 is the (arbitrary) level of the index at time zero. The difficulty then rests only in estimating the relatives.

In general, consider the relative change in rent in month m over month 1, denoted by r_1^m . This " m over 1 relative" can be estimated by

$$\hat{r}_1^m = x_m/x_1. \quad (2.3)$$

However, if one considers matched indices only, the only estimable relatives under the proposed design are the four-month relatives, in other words, those of the form r_{m-4j}^m , $j = 1, 2, 3$, because it is only in these cases that there are common units between the two months. These relatives are estimated by

$$\hat{r}_{m-4j}^m = x_m/x_{m-4j}, \quad (2.4)$$

where the set s of dwellings consists of only those units that report rents at both time m and $m-4j$. Unfortunately, the interest lies in estimating monthly relatives of the form r_{m-1}^m . On the positive side, the rotation scheme ensures that a four-month relative is available every month. It is also assumed that units rotating out of the sample are replaced by equivalent units rotating into the sample. As such, the set s of common dwellings in (2.1) depends on

the time m only and any future reference to it, while implicitly retained, can thus be suppressed in what follows. For a rigorous discussion of these assumptions and the effect on the index if the assumptions fail, the reader is invited to consult Szulc (1983) and Kovar (1984).

In the following paragraphs, five methods of estimating monthly relatives from four-month relatives will be described. Each will be justified intuitively as well as theoretically, and its advantages and disadvantages will be pointed out. The first three methods are derived on a theoretical basis alone while the fourth attempts to exploit the rotation pattern of the survey. All four assume that at least a four month back history of data is available. The last approach takes advantage of prior empirical knowledge: that of high probability of observing one change in rent per year. Methods two and four have been discussed earlier by Kovar (1984).

2.1 Interpolated Index (Additive Index)

One way of estimating the relative r_{m-1}^m is to estimate the previous month's rent, x_{m-1} . This can be accomplished, among other methods, by linearly interpolating the observed rents at time m and $m - 4$, that is, by assuming that the rents increase (decrease) linearly over time. Note that this assumption does not require each individual rent to increase every month by a fixed amount, but merely that the sum of all the rents does. In general, to describe linear interpolation briefly, consider two measurements of the same quantity at two distinct time points, say y_t and y_{t-s} . Suppose that we wish to estimate the value of y at some point between the times $t - s$ and t , say at time $t - u$ ($u < s$). Assuming that the measurements increase linearly in time, y_{t-u} can be estimated from y_t and y_{t-s} by

$$y_{t-u} = (1 - \frac{u}{s})y_t + \frac{u}{s}y_{t-s} \quad (2.5)$$

or in the case at hand, where $s = 4$ and $u = 1$, by

$$y_{t-1} = (3/4)y_t + (1/4)y_{t-4}. \quad (2.6)$$

Thus the previous month's total rent can be estimated by

$$x_{m-1} = (1/4)x_{m-4} + (3/4)x_m \quad (2.7)$$

and consequently, the monthly relative for month m by

$$r_{m-1}^m = \frac{x_m}{x_{m-1}} = \frac{4x_m}{x_{m-4} + 3x_m}. \quad (2.8)$$

The index is then derived by chaining the relatives as in (2.2) above.

Provided that the rents follow the linear interpolation model, that is, provided that we can write the current month's rent as a recursive function of previous months' rents, namely, as

$$x_m = x_{m-1} + d = x_0 + md, \quad (2.9)$$

then it can be shown that the index at time m is given by $I_m = x_m/x_0$, as is desired. In other words, if the data follow the model in (2.9), the index will suffer no time lags. But, of course, if the model were true at all times, the index would be fixed for all time points, based on

any two observations. Since this is clearly not the case, one can at best use (2.8) as an approximation over short periods of time only. In that case, however, if the relationship in (2.9) is not exact, the index at time m will depend on all the rents between time -4 and m . In other words, the index is then susceptible to accumulating various biases over time.

Note that the same index would be derived by assuming that the four-month increment, $x_m - x_{m-4}$, occurred in 4 equal additive steps: $(x_m - x_{m-4})/4$. Since then, the previous month's rent would be estimated by

$$x_{m-1} = x_m - (x_m - x_{m-4})/4, \quad (2.10)$$

which is the same as (2.7); hence the alias: additive index.

2.2 Geometric Index

In this section, in contrast to the above, we will attempt to estimate the relative directly. We first note that

$$\begin{aligned} r_{m-4}^m &= \frac{x_m}{x_{m-4}} = \frac{x_m}{x_{m-1}} \frac{x_{m-1}}{x_{m-2}} \frac{x_{m-2}}{x_{m-3}} \frac{x_{m-3}}{x_{m-4}} \\ &= r_{m-1}^m r_{m-2}^{m-1} r_{m-3}^{m-2} r_{m-4}^{m-3}. \end{aligned} \quad (2.11)$$

We then assume that the four relatives on the right hand side of (2.11) are equal, or equivalently, that the four-month movement is due to four equal movements which act multiplicatively (Kosary *et al.* 1982). Under this assumption, the relationship (2.11) can be written as

$$r_{m-1}^m = (r_{m-4}^m)^{1/4}. \quad (2.12)$$

From (2.2) and (2.3), assuming that there are no sample changes or that units rotating out of the sample are replaced by equivalent units rotating into the sample, the index in month m over the base period zero becomes

$$\begin{aligned} I_m &= I_0 \times r_0^1 \times r_1^2 \times \dots \times r_{m-1}^m \\ &= I_0 \times (r_{-3}^1)^{1/4} \times (r_{-2}^2)^{1/4} \times \dots \times (r_{m-4}^m)^{1/4} \\ &= I_0 \frac{(x_{m-3} x_{m-2} x_{m-1} x_m)^{1/4}}{(x_{-3} x_{-2} x_{-1} x_0)^{1/4}} \end{aligned} \quad (2.13)$$

In other words, the index is a ratio of two geometric averages; hence the name geometric index. We note that at any time, assuming the panels are stationary, the index depends on eight months worth of data only, and thus is independent of any movements between time 0 and $m-4$, though in practice matched sets contributing to each r_{m-4}^m are different, so the cancellation is only theoretical. By contrast the index suffers from one-month to three-month lags and will thus tend to dampen true sudden changes. These changes, however, will be reflected eventually, that is, the index will selfcorrect (Kovar 1984).

As a point of clarification, note also that the relatives in (2.12) can be rewritten as

$$\frac{x_m}{x_{m-1}} = \left[\frac{x_m}{x_{m-4}} \right]^{1/4}$$

or as

$$x_{m-1} = (x_{m-4})^{1/4} (x_m)^{3/4}$$

or finally as

$$\log(x_{m-1}) = (1/4) \log(x_{m-4}) + (3/4) \log(x_m). \quad (2.14)$$

The geometric index is therefore equivalent to an index derived by estimating the previous month's rent by linearly interpolating the logarithms of the observed rents at time m and $m - 4$. (See (2.6) with $y_m = \log x_m$.)

2.3 Incremental Index

Analogous to the above geometric index, here we assume that the four consecutive monthly relative net increments are equal and acting additively. More precisely, we can write r_1^m as

$$r_1^m = 1 + i_1^m$$

where i_1^m is the relative net increment in month m over month 1. To estimate r_{m-1}^m we need therefore i_{m-1}^m . Assuming that the available $i_{m-4}^m = 4i_{m-1}^m$, the relative r_{m-1}^m can be estimated. Namely, we will estimate i_{m-1}^m by

$$i_{m-1}^m = (1/4) i_{m-4}^m = (1/4) (r_{m-4}^m - 1) = (1/4) \left(\frac{x_m}{x_{m-4}} - 1 \right), \quad (2.15)$$

and r_{m-1}^m by

$$r_{m-1}^m = 1 + i_{m-1}^m = \frac{x_m + 3x_{m-4}}{4x_{m-4}}. \quad (2.16)$$

We note that $r_{m-1}^m = x_m/x_{m-1}$ and thus (2.16) can be written as

$$\frac{x_m}{x_{m-1}} = \frac{x_m + 3x_{m-4}}{4x_{m-4}}$$

or as

$$\frac{1}{x_{m-1}} = (1/4) \frac{1}{x_{m-4}} + (3/4) \frac{1}{x_m}. \quad (2.17)$$

In other words, the incremental index corresponds to one which would be derived by estimating the previous month's rent by linearly interpolating the reciprocals of the observed rents at time m and $m - 4$. (See (2.6) with $y_m = x_m^{-1}$.)

As is the case with the interpolated index, the incremental index will be independent of the intermediate observations only under the restrictive condition that the interpolation model be followed. In this case, analogous to (2.9), the model is

$$\frac{1}{x_m} = \frac{1}{x_0} + md. \quad (2.18)$$

However, in most real situations, the chained incremental index will depend on all the data between times -4 and m and therefore will be susceptible to various accumulating biases.

Since all three indices discussed to this point can be described in terms of linear interpolation of various functions of the observed rents, it is also possible to compare them theoretically. It can in fact be shown that the three indices are ordered in magnitude, from smallest to largest in the order of their presentation. That is, in an inflationary situation the interpolated index will always be smaller in absolute value than the geometric index which in turn will always be dominated by the incremental index. The reverse holds true when the trend is downward, that is, when prices are decreasing. As one referee pointed out, this phenomenon can be explained by noting that "the interpolated, geometric and incremental relatives are respectively the weighted arithmetic, geometric, and harmonic means of rent quotations four months apart. The standard relationship between these means explains the behaviour of the estimates in inflationary or deflationary times".

2.4 Carried Index (Arithmetic Index)

The carried index is constructed by taking advantage of the rotating sample at hand. Noting that all units reappear periodically in the sample, we construct the index by simply carrying each unit's rent value forward until a new observation is recorded. In this way all units on the file have a matching previous month's rent and thus the monthly relative, r_{m-1}^m , can be constructed in a straightforward manner. The obvious drawback is that the rent increases (decreases) are not recorded until observed. However, since all changes are eventually recorded, the index will selfcorrect (Kovar 1984) but will suffer from a mixture of one to three-month lags. Just as for the geometric index, sudden (real) changes will be dampened but the carried index will reflect them eventually.

On the technical side, we note that in computing the carried index for any given month one quarter of the observations on the file reflect a four-month movement, whereas three quarters of the observations are carried for one to three months and reflect no change. In fact, in month m we observe x_m and carry x_{m-1} , x_{m-2} and x_{m-3} . Similarly, in month $m-1$ we observe x_{m-1} and carry x_{m-2} , x_{m-3} and x_{m-4} . The monthly relative is therefore given by

$$r_{m-1}^m = \frac{x_m + x_{m-1} + x_{m-2} + x_{m-3}}{x_{m-1} + x_{m-2} + x_{m-3} + x_{m-4}}. \quad (2.19)$$

Chaining the relatives as in (2.2), and assuming again that the samples are stationary, we obtain the index for month m over the base period zero as

$$I_m = I_0 \frac{x_{m-3} + x_{m-2} + x_{m-1} + x_m}{x_{-3} + x_{-2} + x_{-1} + x_0}. \quad (2.20)$$

In other words, the index is a ratio of two arithmetic averages. Analogous to the geometric index, the carried index depends on eight months worth of data only, and thus is independent of the movements between time 0 and $m-4$. As mentioned above, it too suffers from one to three-month lags, and therefore dampens sudden changes.

2.5 Annual Index

Empirical observations suggest that most units change rent once a year. One could therefore argue that yearly relatives are more stable than monthly relatives, since the distribution of individual monthly relatives will necessarily demonstrate two spikes, one around the annual relative and the other at 1. The rotation pattern of the proposed rent pilot (Kovar 1984) ensures that an annual relative be estimable every month, that is that r_{m-12}^m be available. To compute the annual index on a monthly basis, we note that for any chained index the following relationships hold:

$$I_m = r_{m-1}^m I_{m-1} \quad (2.21)$$

and

$$I_m / I_{m-12} = r_{m-12}^m. \quad (2.22)$$

From these relationships we obtain an expression for a monthly relative r_{m-1}^m as

$$r_{m-1}^m = r_{m-12}^m I_{m-12} / I_{m-1}. \quad (2.23)$$

These relatives can then be chained as above to produce an index. Since such a relationship is recursive, we need 12 months worth of indices to be able to "start up". One possibility that exists, is to define the index for the first 12 months, by analogy to the geometric index, as

$$I_k = (r_{k-12}^k)^{k/12}, \quad k = 1, 2, \dots, 12. \quad (2.24)$$

As defined, the annual index is independent of intermediate changes. On the other hand it will be saw-toothed unless individual monthly sample sizes are large. This is due to the fact that consecutive monthly estimates are totally independent. Moreover, it must be noted that the lagging problem will be at least as serious in the case at hand as it is for the indices presented earlier.

3. ADJUSTMENTS

In this section, two adjustment procedures for the above indices will be discussed. First, because the first four indices suffer from one to three month lags, they will smooth out true, sharp peaks. From prior data, it has been observed that rent indices do exhibit sharp rises, in certain cities, with some regularity. To "correct" the smoothed out index, an empirical adjustment will be proposed. By contrast, due to the volatility of the annual index, a smoothing adjustment will also be proposed.

3.1 Empirical Adjustments

It is known, for example, that most rents in Montreal change in July. The first four indices discussed in the previous section would distribute this July change over July, August, September and October. One could however adjust the index in July to reflect a larger change and counter adjust it in the following three months. More precisely, the index could be

multiplied by r^* in the reference month and then by $(r^*)^{-1/3}$ in each of the following three months. Since all the proposed indices are chained indices, in the third month after the reference month the four multipliers will offset each other, leaving no trailing biases. As for the choice of r^* , this will depend on continued empirical observations in each particular city.

It is to be noted that such adjustments must be performed in rare situations only and with great care. It is imperative that the particular situation be monitored, for it is not uncommon for such aberrations to disappear suddenly.

3.2 Smoothing

As a last effort in redeeming a volatile, saw-toothed index, one could consider smoothing it. Like the above adjustments, smoothing should be considered in rare and extreme situations only: in cases where no other alternative exists. The smoothing procedure we consider here involves averaging the index at time m with a linear extrapolation to time m of the smoothed index from time $m - 1$ and $m - 2$. One possible choice of the smoothed index at time m , S_m , is then given by

$$\begin{aligned} S_m &= I_m/2 + (2S_{m-1} - S_{m-2})/2 \\ &= S_{m-1} + (I_m - S_{m-2})/2. \end{aligned} \quad (3.1)$$

Since the smoothing operation basically projects past data into the future, the smoothed index will extend past trends and therefore introduce some lags. Moreover, the method is recursive and consequently could also introduce unwanted biases. Other smoothing methods could be considered, although the utility of smoothing an index that suffers from serious lags is questionable.

4. EMPIRICAL STUDY

The study described in the following paragraphs was initiated in order to test the performance over time of the proposed indices and adjustments. The study provides quantitative information on the ability of the indices to track the true index accurately. It supports the mostly heuristic observations made above and reinforces the theoretical ones.

4.1 The Population

The population of rented dwellings used in this study was designed to duplicate the real situation as closely as possible. For this purpose, the cities, their sizes, and their sample sizes were selected to correspond to those used by the Rent Component of the CPI. Since all real data on rents is available for periods of six months only, the needed thirteen months of data had to be simulated. Eight cities were chosen for this purpose. Some are large, some are small, some have periodic jumps in their indices, but all are CPI index cities and have sufficient amount of rent data available. Moreover, while some of the indices in these cities are strictly increasing, others are both increasing and decreasing.

Only the initial rents of all units (those collected when the unit rotated in) on the CPI rent database for the years 1979 to 1984 inclusive, for the eight cities mentioned above, were

Table 1
Average Sample Sizes (Distinct Units) and the Index at 8401
for Eight Cities Based on the Simulated Population

City	Average Monthly Sample Size	Index at 8401 (8001 = 100)
Halifax	51	144.3
Montreal	268	136.6
Ottawa	35	130.0
Toronto	170	130.4
Winnipeg	105	132.0
Edmonton	112	125.2
Calgary	97	123.5
Vancouver	105	130.5

retrieved. For each unit, twelve additional months worth of data were then simulated using the observed parameters. (This approach is operationally easier than simulating seven months of data in addition to the existing six.) More precisely, for each unit, first a decision was made whether or not a change in rent will occur sometime in the next twelve months. The probability of this event was set to be equal to the observed probability of a rent change in that particular city and year. Then, given that a change was to occur, the appropriate month was selected proportional to the observed incidence of rent changes, again specific to the city and month at hand. The actual amount of the rent change was assumed to be distributed normally with a fixed mean and variance. Robust estimates of these two parameters were obtained from the existing data for each city and each month.

All programming was done in SAS (Statistical Analysis System). The random numbers were generated using the routines RANUNI and RANNOR. The resulting population consists of eight cities and four years of fully rotated data (that is, discarding start up months). The average monthly sample sizes and the value of the simulated index for January 1984 (with Jan 1980 = 100) can be seen for each city in Table 1. The indices, calculated for each of the cities, resemble very closely those observed originally. In the following comparisons, the indices of the simulated population were taken to be the true reference points to be reproduced.

4.2 Comparison of Indices

For the purpose of calculating the indices, it was assumed that of the 13 available observations for each unit, only those for months 1, 5, 9 and 13 were actually observed. All calculations were then based on this (4/13) subsample. The five indices described above were calculated for each city and compared to the true index. All indices are fixed at 100 in January 1980. The empirical adjustment was tested with the Montreal, Halifax and Winnipeg data, for the month of July, January and October respectively. While the results for all the possible combinations of cities and indices are too numerous to include herein, they are available from the author. Some selected highlights will be put forth in the following paragraphs. While not exhaustive, they are hoped to be representative as well as indicative of the situation at hand.

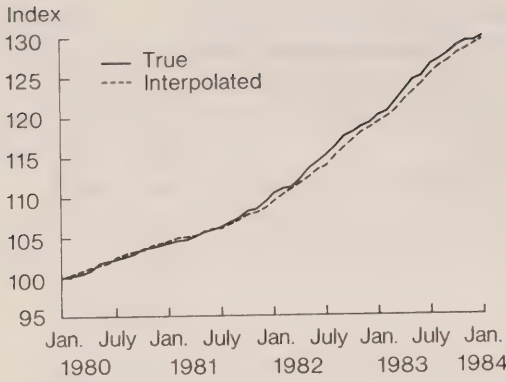


Figure 1. Plot of the True Index and the Interpolated Index for the City of Ottawa

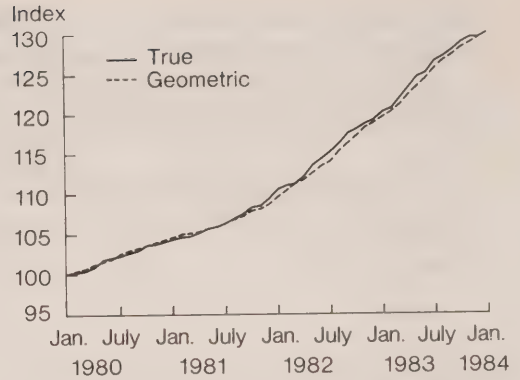


Figure 2. Plot of the True Index and the Geometric Index for the City of Ottawa

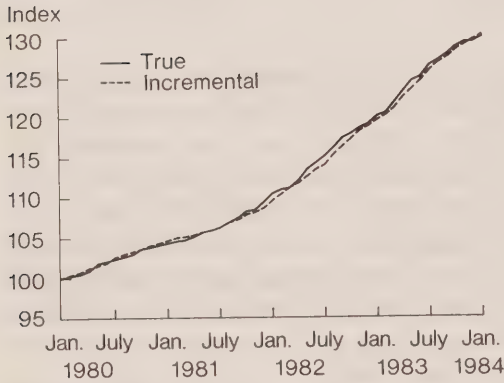


Figure 3. Plot of the True Index and the Incremental Index for the City of Ottawa

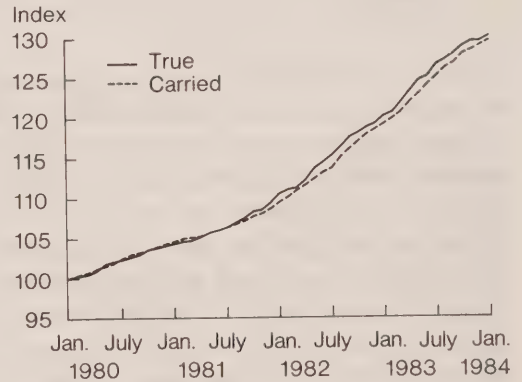


Figure 4. Plot of the True Index and the Carried Index for the City of Ottawa

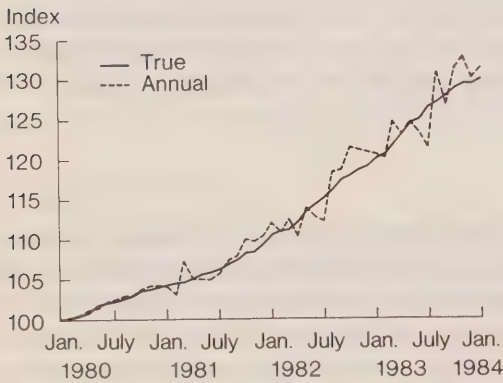


Figure 5. Plot of the True Index and the Annual Index for the City of Ottawa

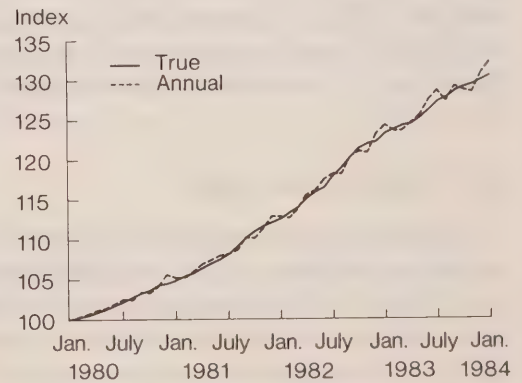


Figure 6. Plot of the True Index and the Annual Index for the City of Toronto

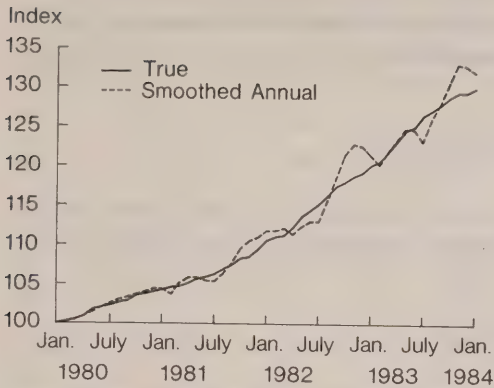


Figure 7. Plot of the True Index and the Smoothed Annual Index for the City of Ottawa

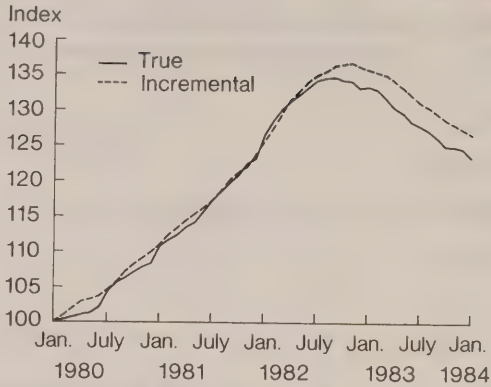


Figure 8. Plot of the True Index and the Incremental Index for the City of Calgary

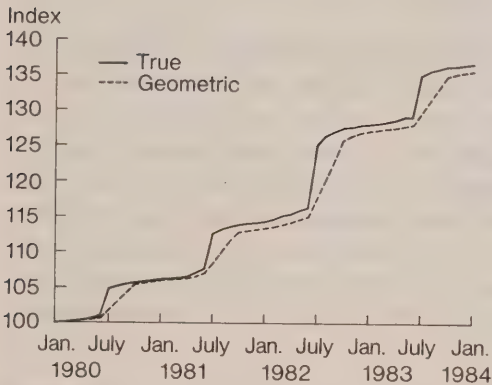


Figure 9. Plot of the True Index and the Geometric Index for the City of Montreal

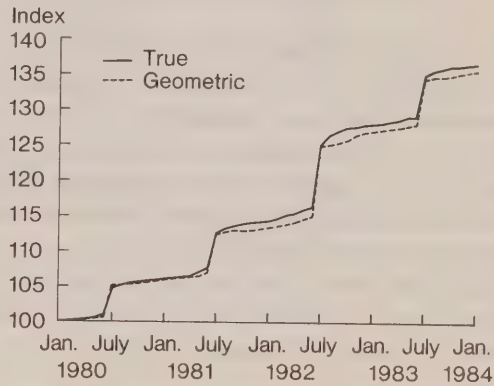


Figure 10. Plot of the True Index and the Adjusted Geometric Index for the City of Montreal

As can be seen in Figures 1-5, all five indices track the true index reasonably well, even in the case of small sample sizes such as in the city of Ottawa. As expected, the first four indices show some lags, those being more pronounced in the carried and interpolated index. (Note that the lagging problem could likely be accentuated by generating the population with exponentially increasing prices). Not surprisingly, the annual index is rather volatile. For cities with large sample sizes however, (e.g. Toronto), the annual index performs well (see Figure 6). While the smoothing adjustment of Section 3.2 does indeed smooth the index, the results are less than satisfactory as can be seen in Figure 7 (c.f. Figure 5). Perhaps a larger number of points should be used for the extrapolation but then the lagging problem would be even more pronounced. Figure 8 further demonstrates how sudden unexpected changes in trends are reported with a delay. However, expected jumps in the index (as in July in Montreal, Figure 9) can be adjusted successfully using the adjustment procedure of Section 3.1 (Figure 10).

Table 2
Mean Square Errors of Five Indices in Eight Cities

City	Interpolated		Geometric		Incremental		Carried		Annual	
Halifax	30*	(3)	19*	(2)	12*	(1)	48	(4)	74	(5)
Montreal	48*	(3)	24*	(2)	9*	(1)	160	(5)	82	(4)
Ottawa	17	(3)	12	(2)	8	(1)	22	(4)	95	(5)
Toronto	36	(4)	27	(3)	20	(2)	29	(5)	13	(1)
Winnipeg	27*	(3)	17*	(2)	10*	(1)	66	(5)	41	(4)
Edmonton	46	(1)	64	(4)	88	(5)	55	(3)	50	(2)
Calgary	56	(2)	81	(4)	121	(5)	64	(3)	46	(1)
Vancouver	70	(5)	53	(2)	39	(1)	64	(4)	60	(3)

Note: 1. Bracketed figures indicate ranking within cities.

2. Starred figures are results of adjusted indices as per Section 3.1.

Mean square errors of the five indices away from the true index have been calculated for each city (Table 2). The three interpolation based indices (interpolated, geometric and incremental) have been adjusted for the cities of Montreal, Halifax and Winnipeg. Table 2 also presents the rankings (from smallest to largest) of the mean square errors of the five indices within each city. The carried and the annual index tend to perform the worst. The three interpolation-based indices perform relatively alike. In general, in cities where the index is climbing consistently, the performance of these three indices worsens in the order: incremental, geometric, interpolated. The order is reversed in cities where sharp decreases in the index have been observed. It is unlikely, however, that the strategies could be interchanged based on observed behaviours only.

5. SUMMARY

Both the theoretical as well as the empirical observations suggest that the yearly index is too volatile in cities where sample sizes are not large enough. Smoothing, at least of the type described, has proven fruitless. For this reason the annual index should be reserved only for those rare cases where sample sizes permit. On the other hand, the annual index could be used in conjunction with one of the more stable four-month indices to produce a composite estimate analogous to that proposed by Kosary *et al.* (1982). However, empirical observations would be needed to determine the appropriate weights to be used in averaging the two indices.

By contrast, the carried, and to some degree, the interpolated index tend to be too smooth. That is they tend to smooth out all peaks in addition to demonstrating a one or two (index) point lag. While the incremental and geometric indices are not entirely free of these lags, they tend to track the true index a little more closely. The incremental index performs the best overall, however, because of the mathematical "cleanliness" of the geometric index (i.e. its theoretical independence of its history and its correspondence to the chaining structure), it is the latter that is recommended here. In other words, the geometric index does not retain terms that could cause biases in the long run.

It is also apparent that whenever possible, prior knowledge can be used to improve the index. Empirical adjustments as described in Section 3.1 can be useful, provided that they are well founded. If their use is contemplated, it is imperative that the empirical knowledge that leads to their application be monitored and its continued existence verified.

ACKNOWLEDGEMENT

Thanks are due to Mr. George Sampson for all of his help and to the referees and the editorial staff for their constructive comments.

REFERENCES

- DOLSON, D.D. (1982). Rent status survey: Analysis. Technical Report, Statistics Canada.
- KOSARY, C.L., BRANSCOME, J.M. and SOMMERS, J.P. (1982). Evaluating alternatives to the rent estimator. Technical Report, Bureau of Labor Statistics.
- KOVAR, J. (1984). Note on calculating the rent index. Technical Report, Statistics Canada.
- SZULC, B. (1983). Linking price index numbers. Technical Report, Statistics Canada.

Regression Analysis Using Survey Data with Endogenous Design

ARIE TEN CATE¹

ABSTRACT

This paper discusses the influence of the sampling design on the estimation of a linear regression model. Particularly, sampling designs will be discussed which are dependent on the values of the endogenous variable in the population: endogenous (or "informative") designs. A consistent estimator of the regression coefficients is given. Its variance is the sum of a sampling design component and a disturbance term component. Also, model-free regression is briefly discussed. The model-free regression estimator is the same as the model estimator in the case of an endogenous design.

KEY WORDS: Regression; Survey sampling; Endogenous design.

1. INTRODUCTION

The heart of any statistical model is the assumption that the value of one or more variables is generated by drawing from some probability distribution; for example, a regression model with normally distributed disturbances. In this paper a finite set of elements which behave according to such a model will be considered. This set is called the population. Next, a sample is drawn from this population, without replacement. The subject of this paper is the influence of the sampling design on the estimation of the parameters of the model. This influence depends mainly on whether the design is exogenous or endogenous with respect to the model. In the case of an endogenous (or "informative") design, the sampling probabilities depend on the value of the endogenous ("dependent") variables. Then, the design should not be ignored in the estimation of the model parameters. The nature of the problem is indicated in Figure 1, where a stratified sampling design is shown. There are 3 strata, defined in the endogenous variable of a regression model. The middle stratum has a higher sampling fraction than the other two. The diagram shows that the slope of the regression line estimated using the sampled data points only, is biased downwards if one ignores the design. This bias does not vanish in large samples. This can be seen in an intuitive manner by imagining that every white and black dot in Figure 1 denotes a large number of identical data points. Even if this large number tends to infinity, the slope of the estimated regression line will be biased downwards, because the shape of the scatter will remain the same.

There is a rapidly growing body of literature on the application of regression techniques in finite population sampling. This literature deals with a variety of problems. One problem is, how to use regression techniques in order to estimate a finite population total. Another problem concerns the estimation of population parameters such as $\Sigma xy / \Sigma x^2$, where the summation runs over all elements of the finite population. Reviews of the literature about these problems are given by Nathan (1981) and Smith (1981). A third problem is the estimation of the parameters of a regression model, using a sample from a finite population. This problem can be solved relatively easily in the case of a exogenous design. See Porter (1973, Section 1.2), DuMouchel and Duncan (1983), and textbooks such as Cramer (1971, p. 143). Texts

¹ Arie ten Cate, Central Planning Bureau, 2585 JR 's-Gravenhage, Van Stolkweg 14, The Hague, The Netherlands.

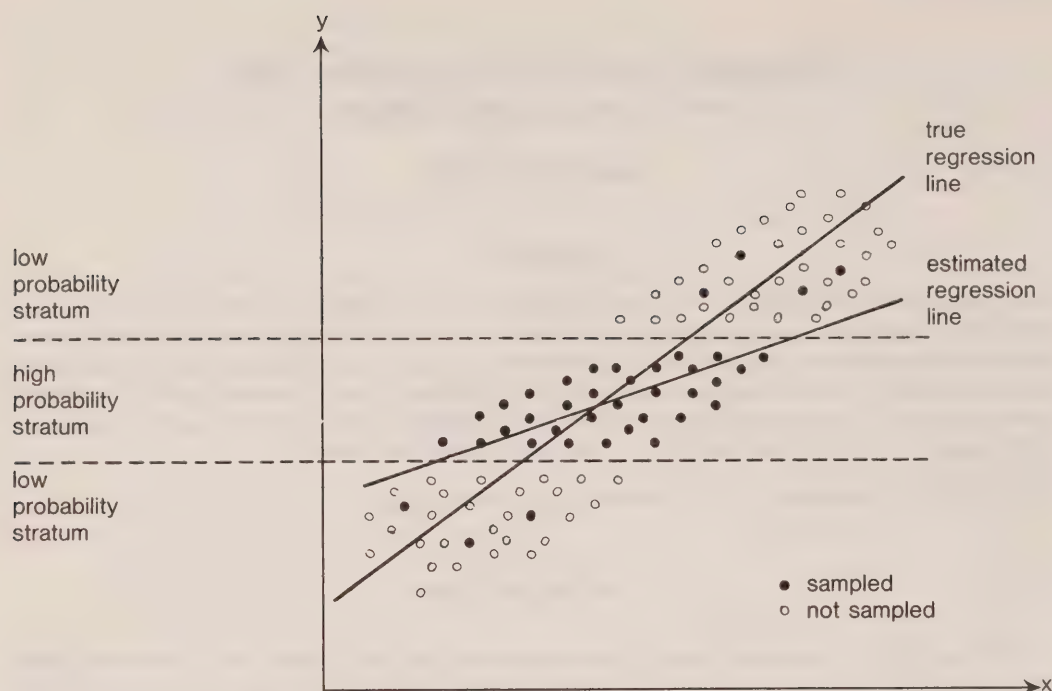


Figure 1. The Effect of Endogenous Stratification on the Estimated Regression Line

such as Kmenta (1971, Section 8.3) and Johnston (1972, Section 9.2) discuss the closely related topic of stochastic regressors. See also White (1980a) for non-linear regression. Our topic, regression analysis with endogenous design, is more complicated. Hausman and Wise (1981) discuss stratified endogenous designs in a very simple case: two strata and a regression model consisting of a constant term only. Jewell (1985) gives some iterative estimators for the case of endogenous stratification.

Regression analysis with endogenous design is related to the problem of endogenous non-response in regression analysis (see Heckman (1979)). However, we have a lesser problem here, since the probabilities involved in the sampling process are assumed to be known: they constitute the chosen design. On the other hand, as we shall see in Subsection 6.1, variance estimation with an endogenous design is in general rather difficult.

Regression analysis with endogenous design may be compared with logit analysis with endogenous design, also called logit analysis with choice based sampling or case-control sampling. See Manski and McFadden (1981, Chapters 1 and 2) and Breslow and Day (1980, Section 6.3).

The contents of the rest of the paper are as follows. In Sections 2 and 3 the main theorems are given. These theorems give a consistent estimator of the parameters of a linear regression model, using a sample with an endogenous design. Consistency is defined here in a similar way as in the discussion of the bias in the example above, though slightly more subtle: the x -values are replicated a large number of times and the y -values behave according to the regression model. In Sections 4 and 5 the variance of the estimator of the regression coefficients is studied. Section 6 discusses the estimation of this variance. Section 7 deals with model-free regression, Section 8 discusses the various motives for weighted regression and finally, Section 9 concludes the paper.

2. THE MODEL, THE SAMPLE AND A REGRESSION ESTIMATOR

In this section the asymptotic properties of an estimator of a regression model are studied within the framework of finite population sampling without replacement. Asymptotic theory for samples drawn without replacement from a finite population may seem a contradiction since such a sample must be bounded. This contradiction is solved by increasing both the population size and the sample size, without bound, at the same rate. The dependence between the inclusions of population elements in the sample constitutes another problem, especially in the case of complex sampling designs. Here we use an idea of Brewer (1979). In Brewer's system, limit theorems on sequences of independent variables can be used, while the results may still be applied to complex designs. Basically, this system consists of the replica idea already introduced informally above. This replica idea will be used extensively throughout the rest of this paper. For another approach, see Robinson (1982).

First, the structure of the population and the model are given. Consider a finite set of N_0 elements. Each element has r real-valued exogenous non-stochastic characteristics, together forming an $(N_0 \times r)$ -matrix X_0 . One of the fundamental assumptions of this paper is the following. The population consists of K replicas of this set of N_0 elements, having $N \equiv KN_0$ elements. Its matrix of exogenous variables is X , with

$$X = \iota_K \otimes X_0. \quad (1)$$

Here, ι_K is the K -vector with all elements equal to unity and \otimes denotes the Kronecker matrix product. Asymptotic results will be derived by allowing K to tend to infinity.

The model assumptions describe the standard linear model. Each of the N elements of the population has a score on a stochastic, endogenous, variable. Together they form an N -vector y . It is assumed that

$$E_\xi(y) = X\beta \quad (2)$$

for some fixed, unknown r -vector β . E_ξ denotes the expectation over all $y \in R^N$. Next we define

$$\varepsilon = y - X\beta. \quad (3)$$

It is assumed that the N elements of ε are i.i.d. It follows from (2) that all elements of ε have expectation zero. Their variance is σ^2 , that is,

$$E_\xi(\varepsilon\varepsilon') = \sigma^2 I. \quad (4)$$

Sampling is done without replacement here, as is common practice. The sample is described by a diagonal $(N \times N)$ -matrix T , such that

$$t_{ii} = \begin{cases} 1 & \text{if population element } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}$$

for all $i = 1, \dots, N$. Obviously, T is idempotent. The sample space S is the set of all such matrices T . This set is finite. The sampling design is some probability distribution over the elements of the sample space S . The sampling design is endogenous here, meaning that it depends on y . Hence, the sampling design itself is stochastic. (A design which does not depend on y is called exogenous, or uninformative.) Let T be partitioned in a square $K \times K$ array of $(N_0 \times N_0)$ blocks. Let T_k be the k -th diagonal block, related to the k -th replica. Similarly, let y be partitioned in K N_0 -vectors, such that $y' = (y'_1, y'_2, \dots, y'_k, \dots, y'_K)$. It is assumed that the sampling design depends on y in the following sense: the K pairs $(T_1, y_1), \dots, (T_K, y_K)$ are i.i.d.

The expectation over all elements of S , conditional on y (or ϵ), plays an important role in this paper. It is denoted by E_p . Then we define

$$\Pi \equiv E_p(T). \tag{5}$$

It is assumed that Π is known. The diagonal elements of Π are called inclusion probabilities: the probabilities that the population elements are included in the sample. The matrix Π is partitioned in a square $K \times K$ array of $(N_0 \times N_0)$ blocks. Let Π_k be the k -th diagonal block, related to the k -th replica. Note that each Π_k is stochastic because it depends on y_k . By the above assumption, the Π_1, \dots, Π_K are i.i.d. The dependence of the Π_k on y is denoted by a function F , such that

$$\Pi_k = F(y_k) \tag{6}$$

for all $k = 1, \dots, K$. It is assumed that $F(y_k)$ is non-singular for every y_k . In other words, the inclusion probabilities are always positive.

This framework and Brewer's (1979) differ in somewhat. Brewer has no endogenous variables and therefore all his Π_k are nonstochastic and equal. One may also compare this approach with the idea of "constant in repeated samples" in the econometric literature; see e.g. Theil (1971, p. 364).

The stage is now set for the estimation of β . The stochastic properties of estimators will be considered over all pairs $(y, T) \in (R^N \times S)$. The corresponding expectation will be denoted by $E_\epsilon E_p$. We shall consider a generalized least square estimator of β , say $\hat{\beta}$, with weights equal to the square roots of the inclusion probabilities, as follows,

$$\begin{aligned} \hat{\beta} &\equiv [(\Pi^{-1/2}X)'T(\Pi^{-1/2}X)]^{-1}(\Pi^{-1/2}X)'T(\Pi^{-1/2}y) \\ &= (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}Ty. \end{aligned} \tag{7}$$

Recall that the matrix Π is known. Note that X and y relate to the population, but T effectuates summation over the sampled elements. As an alternative to considering $\hat{\beta}$ as a generalized least squares estimator, assume that all elements of Π^{-1} are integer numbers. Then, if each observation i in the sample is copied π_{ii}^{-1} times, $\hat{\beta}$ is the ordinary least squares estimator applied to this inflated sample. In this view, no square roots of the probabilities are involved. See also Hausman and Wise (1981, p. 373). The main theorem of this paper is:

Theorem 1. Under the assumptions made above ((1), (2) and the distribution of ϵ and T), the generalized least squares estimator $\hat{\beta}$, defined in equation (7) is consistent for $K \rightarrow \infty$.

The rest of the section is devoted to the proof of this theorem. The following lemma will be used in this proof and the proof of subsequent theorems.

Lemma 1. Consider an N -vector z , such that $z = \iota_k \otimes z_0$, where z_0 is some fixed N_0 -vector. Consider also an N -vector η , partitioned such that $\eta' = (\eta'_1, \eta'_2, \dots, \eta'_K)$. Each η_k has N_0 elements. Assume that each η_k is a function of X_0 , β and ε_k , all functions being the same. Then

$$\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} z' \Pi^{-1} T \eta \right) = z'_0 E_{\xi}(\eta_0), \tag{8}$$

where $E_{\xi}(\eta_0)$ is the expectation of any η_k , being equal for all k .

Proof of lemma 1: Consider the expectation of $\Pi_k^{-1} T_k \eta_k$:

$$E_{\xi} E_p (\Pi_k^{-1} T_k \eta_k) = E_{\xi} [\Pi_k^{-1} E_p (T_k) \eta_k] = E_{\xi}(\eta_k), \tag{9}$$

for all k . Since the distribution of η_k is the same for each k , one may write

$$E_{\xi} E_p (\Pi_k^{-1} T_k \eta_k) = E_{\xi}(\eta_0) \tag{10}$$

for all k . Also, the K vectors $z'_0 \Pi^{-1} T_k \eta_k$ are i.i.d. Thus, Khintchine's theorem applies as follows,

$$\begin{aligned} \text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} z' \Pi^{-1} T \eta \right) &= \text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \sum_k z'_0 \Pi_k^{-1} T_k \eta_k \right) = E_{\xi} E_p (z'_0 \Pi_1^{-1} T_1 \eta_1) \\ &= z'_0 E_{\xi} E_p (\Pi_1^{-1} T_1 \eta_1). \end{aligned} \tag{11}$$

Substitution of (10) in (11) gives the lemma. The proof of theorem 1 is now straightforward.

Proof of theorem 1: The generalized least squares estimator of the theorem can be written as

$$\hat{\beta} = (X' \Pi^{-1} T X)^{-1} X' \Pi^{-1} T y = \beta + (X' \Pi^{-1} T X)^{-1} X' \Pi^{-1} T \varepsilon. \tag{12}$$

Thus,

$$\begin{aligned} \text{plim}_{K \rightarrow \infty} \hat{\beta} &= \beta + \left[\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T X \right) \right]^{-1} \text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T \varepsilon \right) \\ &= \beta + (X'_0 X_0)^{-1} X'_0 0 = \beta. \end{aligned} \tag{13}$$

The expression $X'_0 X_0$ is formed by repeated application of lemma 1, substituting the columns of X for both z and η . Notice that $E_{\xi}(X_0) = X_0$ since X_0 is a constant. The expression $X'_0 0$ is formed by repeated application of lemma 1, substituting the columns of X for z and ε for η .

3. THE ESTIMATION OF THE DISTURBANCE VARIANCE

The regression model described in Section 2 has two parameters: β and σ^2 . Theorem 1 considered estimation of β ; in this section the estimation of σ^2 will be considered. The result of this section is given in the following theorem.

Theorem 2. The disturbance variance σ^2 is estimated consistently by the weighted sample variance of the residuals of y if these weights are equal to the inverse of the square root of the inclusion probabilities.

Proof: The variance estimator of the theorem is

$$\hat{\sigma}^2 = (\iota'_N \Pi^{-1} T \iota_N)^{-1} \tilde{e}' \tilde{e} \quad (14)$$

with

$$\tilde{e} \equiv \Pi^{-1/2} T(y - X\hat{\beta}). \quad (15)$$

Let

$$\tilde{y} \equiv \Pi^{-1/2} Ty, \quad (16)$$

$$\tilde{X} \equiv \Pi^{-1/2} TX, \quad (17)$$

and

$$\tilde{\varepsilon} \equiv \Pi^{-1/2} T\varepsilon. \quad (18)$$

Then

$$\tilde{e} = \tilde{y} - \tilde{X}\hat{\beta} = \tilde{y} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \quad (19)$$

and

$$\begin{aligned} \tilde{e}'\tilde{e} &= \tilde{y}' [I_N - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}']\tilde{y} = (\tilde{X}\hat{\beta} + \tilde{\varepsilon})' [I_N - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'] (\tilde{X}\hat{\beta} + \tilde{\varepsilon}) \\ &= \tilde{\varepsilon}'\tilde{\varepsilon} - \tilde{\varepsilon}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\varepsilon}. \end{aligned} \quad (20)$$

The first term in the right-hand side (RHS) of (20) converges in probability as follows

$$\begin{aligned} \text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \tilde{\varepsilon}' \tilde{\varepsilon} \right) &= \text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \varepsilon' \Pi^{-1} T \varepsilon \right) = \text{plim}_{K \rightarrow \infty} \left[\frac{1}{K} \iota'_N \Pi^{-1} T \text{diag}(\varepsilon) \varepsilon \right] \\ &= \iota'_{N_0} (\sigma^2 \iota_{N_0}) = N_0 \sigma^2. \end{aligned} \quad (21)$$

Here, $\text{diag}(\varepsilon)$ indicates the diagonal matrix with ε as the diagonal. Lemma 1 has been applied with ι_N substituted for z and $\text{diag}(\varepsilon)\varepsilon$ for η , using model equation (4). Next, consider the second term in the RHS of (20).

$$\begin{aligned}
 & \text{plim}_{K \rightarrow \infty} \left[\frac{1}{K} \tilde{\varepsilon}' \tilde{X} (\tilde{X}' \tilde{X})^{-1} \tilde{X}' \tilde{\varepsilon} \right] \\
 &= \left[\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \tilde{X}' \tilde{\varepsilon} \right) \right]' \left[\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \tilde{X}' \tilde{X} \right) \right]^{-1} \text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \tilde{X}' \tilde{\varepsilon} \right) \\
 &= \left[\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T \varepsilon \right) \right]' \left[\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T X \right) \right]^{-1} \text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T \varepsilon \right) \\
 &= 0' (X_0' X_0)^{-1} 0 = 0.
 \end{aligned} \tag{22}$$

In the derivation of (22), use has been made of lemma 1 in the same manner as in the derivation of (13). The combination of (20), (21) and (22) gives

$$\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \tilde{\varepsilon}' \tilde{\varepsilon} \right) = N_0 \sigma^2. \tag{23}$$

Finally, lemma 1 is applied to the first factor in (14), with ι_N substituted both for z and η . This gives

$$\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \iota_N' \Pi^{-1} T \iota_N \right) = N_0. \tag{24}$$

With (23) and (24) we have

$$\text{plim}_{K \rightarrow \infty} (\hat{\sigma}^2) = \sigma^2, \tag{25}$$

which proves the theorem. Finally it may be useful to note, as a corollary of (23), that

$$\left(\frac{1}{N} \right) \tilde{\varepsilon}' \tilde{\varepsilon} \tag{26}$$

is also a consistent estimator of σ^2 .

4. THE VARIANCE OF $\hat{\beta}$

In this section the asymptotic variance of the estimator $\hat{\beta}$ is given.
Theorem 3. The asymptotic variance of $\hat{\beta}$ is given by

$$\text{Var}(\hat{\beta}) = (X' X)^{-1} X' V X (X' X)^{-1}, \tag{27}$$

with

$$V \equiv E_{\xi} [\text{diag}(\varepsilon) \Pi^{-1} P \Pi^{-1} \text{diag}(\varepsilon)], \tag{28}$$

and

$$P \equiv E_p(T\iota\iota'T). \quad (29)$$

The elements of P are the so-called second order inclusion probabilities: the probability for any pair of elements of the population of being included in the sample. The diagonal of P is equal to the diagonal of Π . The rest of this section is devoted to a proof of this theorem.

Proof: Consider the asymptotic distribution for $K \rightarrow \infty$ of

$$\begin{aligned} K^{1/2}(\hat{\beta} - \beta) &= K^{1/2}[(X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}Ty - \beta] \\ &= K^{1/2}(X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}T\varepsilon. \end{aligned} \quad (30)$$

Since

$$\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X'\Pi^{-1}TX \right) = X_0'X_0, \quad (31)$$

the asymptotic distribution of $K^{1/2}(\hat{\beta} - \beta)$ is equal to the asymptotic distribution of δ , with

$$\delta \equiv K^{-1/2}(X_0'X_0)^{-1}X_0'\Pi^{-1}T\varepsilon = K^{-1/2}(X_0'X_0)^{-1} \sum_k X_0'\Pi_k^{-1}T_k\varepsilon_k = K^{-1/2} \sum_k \delta_k, \quad (32)$$

and

$$\delta_k \equiv (X_0'X_0)^{-1}X_0'\Pi_k^{-1}T_k\varepsilon_k, \quad (33)$$

for all $k = 1, \dots, K$. (See e.g. Rao (1973), p. 122). Since the vector δ_k ($k = 1, \dots, K$) are i.i.d. and also

$$\begin{aligned} E_\xi E_p(\delta_k) &= (X_0'X_0)^{-1}X_0'E_\xi E_p(\Pi_k^{-1}T_k\varepsilon_k) \\ &= (X_0'X_0)^{-1}X_0'E_\xi [\Pi_k^{-1}E_p(T_k)\varepsilon_k] \\ &= (X_0'X_0)^{-1}X_0'E_\xi(\varepsilon_k) = 0, \end{aligned} \quad (34)$$

the variance of δ , say $\text{Var}(\delta)$, is equal for all K and also equal to the variance of the asymptotic distribution of δ for $K \rightarrow \infty$. This variance can be written as

$$\text{Var}(\delta) = E_\xi E_p(\delta_k \delta_k') \quad (35)$$

for any $k \in \{1, \dots, K\}$. Since the vectors δ_k are i.i.d. this may be rewritten as

$$\begin{aligned}
\text{Var}(\delta) &= \frac{1}{K} \sum_k E_{\xi} E_p(\delta_k \delta_k') \\
&= \frac{1}{K} (X_0' X_0)^{-1} \left[E_{\xi} E_p \left(\sum_k X_0' \Pi_k^{-1} T_k \varepsilon_k \varepsilon_k' T_k \Pi_k^{-1} X_0 \right) \right] (X_0' X_0)^{-1} \\
&= K(X'X)^{-1} [E_{\xi} E_p(X' \Pi^{-1} T \varepsilon \varepsilon' T \Pi^{-1} X)] (X'X)^{-1} \\
&= K(X'X)^{-1} X' \{E_{\xi} E_p[\text{diag}(\varepsilon) \Pi^{-1} T \iota \iota' T \Pi^{-1} \text{diag}(\varepsilon)]\} X(X'X)^{-1} \\
&= K(X'X)^{-1} X' \{E_{\xi} [\text{diag}(\varepsilon) \Pi^{-1} E_p(T \iota \iota' T) \Pi^{-1} \text{diag}(\varepsilon)]\} X(X'X)^{-1}. \quad (36)
\end{aligned}$$

Division of $\text{Var}(\delta)$ by K gives $\text{Var}(\hat{\beta})$ and completes the proof.

5. A DECOMPOSITION OF $\text{VAR}(\hat{\beta})$

The variance formula (27) can be rewritten as

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1} + (X'X)^{-1} X' V^* X (X'X)^{-1} \quad (37)$$

with

$$V^* \equiv E_{\xi} [\text{diag}(\varepsilon) (\Pi^{-1} P \Pi^{-1} - \iota \iota') \text{diag}(\varepsilon)], \quad (38)$$

using (4). The first term in the RHS of (37) might reasonably be called the ξ -component of the variance of $\hat{\beta}$. This component would contain all the variance of $\hat{\beta}$ if the whole population was sampled. It is entirely due to the variation in the disturbance ε and it is the familiar expression for that case. The second term in the RHS of (37) might be called the p -component of the variance of $\hat{\beta}$. This component contains the matrices Π and P , which describe the sampling design. This component looks like the variance formula of the estimator of a total or average of a finite population. The theory of such estimators will be discussed briefly in the rest of this section, as an aid in the interpretation of the p -component of $\text{Var}(\hat{\beta})$.

Consider a finite population of N elements. (No replica structure is assumed here). Each element of this population has a score on some real non-stochastic variable, collected in an N -vector x . From this population a sample without replacement is taken. The sample is described by the diagonal matrix T , as before. Also as before,

$$\Pi \equiv E_p(T) \quad (39)$$

and

$$P \equiv E_p(T \iota \iota' T), \quad (40)$$

the first order and second order inclusion probabilities, respectively. There is no regression model here, so Π and P are fixed known matrices. Horvitz and Thompson (1952) suggested to estimate the population total $X' \iota$ by

$$\hat{X} = x' \Pi^{-1} T \iota \quad (41)$$

Obviously this is an unbiased estimator, in view of (39). The variance of \hat{X} is

$$\begin{aligned}\text{Var}(\hat{X}) &= E_p(\hat{X}^2) - [E_p(\hat{X})]^2 = E_p(x' \Pi^{-1} T \iota \iota' T \Pi^{-1} x) - x' \iota \iota' x \\ &= x' (\Pi^{-1} P \Pi^{-1} - \iota \iota') x.\end{aligned}\quad (42)$$

The last member of equation (42) is the variance formula of the Horvitz-Thompson estimator, which can be found in textbooks on sampling, such as Cochran (1977), though usually not in matrix format. The expression in parentheses in the last member of (42) is equal to the expression in parentheses in (38), the definition of V^* . The latter is contained in the formula of the p -component of $\text{Var}(\hat{\beta})$. Thus, the diagonal elements of the p -component of the variance matrix $\text{Var}(\hat{\beta})$ can be considered as the ξ -expectation of the p -variance of the Horvitz-Thompson estimator of the row totals of $(X'X)^{-1} X' \text{diag}(\epsilon)$. These totals are the elements of the vector $(X'X)^{-1} X' \epsilon$.

6. THE ESTIMATION OF $\text{Var}(\hat{\beta})$

6.1 The General Case

In this section the estimation of the asymptotic variance $\text{Var}(\hat{\beta})$ is considered. Consistent estimation of $\text{Var}(\hat{\beta})$ is rather difficult, since this requires knowledge of the relationship F between y and the sampling design, as it appears in the matrix V . In practice, only the sampling design for the actual values of y will be known. In general, it is difficult to tell from this design only, what the design would be like if y took on different values. In a sense not only a regression model is involved, but also a model of the designer himself!

For the moment we assume that the function F is known, and therefore V is a known function of X and the parameters of the model. (See Subsection 6.2 for a special case). This is expressed as follows.

$$V = V(\beta, \sigma^2; X), \quad (43)$$

It is assumed that $V(\beta, \sigma^2; X)$ is a continuous function. For the sake of brevity, \hat{V} is defined as

$$\hat{V} \equiv V(\hat{\beta}, \hat{\sigma}^2; X), \quad (44)$$

where $\hat{\beta}$ and $\hat{\sigma}^2$ are consistent estimators of β and σ^2 respectively. The rest of this subsection gives a theorem on consistent variance estimation, and its proof. Consistent estimation of $\text{Var}(\hat{\beta})$ by $\hat{\text{var}}(\hat{\beta})$ is interpreted here as follows:

$$\text{plim}_{K \rightarrow \infty} K \hat{\text{var}}(\hat{\beta}) = \lim_{K \rightarrow \infty} K \text{Var}(\hat{\beta}). \quad (45)$$

Theorem 4. Under the assumptions made above, the asymptotic variance $\text{Var}(\hat{\beta})$ is estimated consistently by

$$\hat{\text{var}}(\hat{\beta}) = (X' \Pi^{-1} T X)^{-1} X' T \left(\frac{\hat{V}}{P} \right) T X (X' \Pi^{-1} T X)^{-1}, \quad (46)$$

where (\hat{V}/P) denotes the matrix consisting of the elements of \hat{V} divided by the corresponding elements of P .

Proof: First the structure of V will be considered. Let V be partitioned in a square $K \times K$ array of $(N_0 \times N_0)$ blocks. The (k, r) -th off-diagonal block of V is equal to

$$\begin{aligned} E_{\xi} [\text{diag}(\varepsilon_k) \Pi_k^{-1} E_p(T_k \iota \iota' T_r) \Pi_r^{-1} \text{diag}(\varepsilon_r)] \\ = E_{\xi} [\text{diag}(\varepsilon_k) \Pi_k^{-1} E_p(T_k) \iota \iota' E_p(T_r) \Pi_r^{-1} \text{diag}(\varepsilon_r)] \\ = E_{\xi} (\varepsilon_k \varepsilon_r') = 0, \end{aligned} \quad (47)$$

using the assumed replica structure of the population and the sampling design. The diagonal blocks of V are identical and depend on X_0 . Thus, $V(\beta, \sigma^2; X)$ can be written as

$$V(\beta, \sigma^2; X) = I_K \otimes V_0(\beta, \sigma^2; X_0), \quad (48)$$

where $V_0(\beta, \sigma^2; X_0)$ is an $N_0 \times N_0$ matrix function. Together with (1), equation (48) can be used to rewrite $K\text{Var}(\hat{\beta})$ as follows.

$$K\text{Var}(\hat{\beta}) = (X_0' X_0)^{-1} X_0' V_0 X_0 (X_0' X_0)^{-1}, \quad (49)$$

where V_0 denotes $V_0(\beta, \sigma^2; X_0)$. The RHS of (49) is independent of K and therefore equal to its limit as K tends to infinity. Next, the LHS of (45) is considered.

$$K\hat{\text{var}}(\hat{\beta}) = \left(\frac{1}{K} X' \Pi^{-1} T X \right)^{-1} \left[\frac{1}{K} X' T \left(\frac{\hat{V}}{P} \right) T X \right] \left(\frac{1}{K} X' \Pi^{-1} T X \right)^{-1}. \quad (50)$$

Earlier, in the derivation of (13) and (22), use has already been made of

$$\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T X \right) = X_0' X_0. \quad (51)$$

It follows from the assumption that $V(\beta, \sigma^2; X)$ is a continuous function, that

$$\text{plim}_{K \rightarrow \infty} \hat{V}_0 = V_0, \quad (52)$$

where \hat{V}_0 denotes $V_0(\hat{\beta}, \hat{\sigma}^2; X_0)$. Using (1), (48) and (52) gives

$$\begin{aligned} \text{plim}_{K \rightarrow \infty} \frac{1}{K} X' T \left(\frac{\hat{V}}{P} \right) T X &= \text{plim}_{K \rightarrow \infty} \frac{1}{K} \sum_k \left[X_0' T_k \left(\frac{\hat{V}_0}{P_0} \right) T_k X_0 \right] \\ &= \text{plim}_{K \rightarrow \infty} \frac{1}{K} \sum_k \left[X_0' T_k \left(\frac{V_0}{P_0} \right) T_k X_0 \right] = X_0' V_0 X_0. \end{aligned} \quad (53)$$

Here P_0 denotes $E_p (T_k u' T_k)$, which is the same for all $k = 1, \dots, K$. The last equality sign results from the application of Khintchine's theorem, since the terms in the second summation over k in (53) are i.d.d. with p -expectation equal to $X_0' V_0 X_0$. Finally, the combination of (50), (51) and (53) gives

$$\text{plim}_{K \rightarrow \infty} K \text{var}(\hat{\beta}) = (X_0' X_0)^{-1} X_0' V_0 X_0 (X_0' X_0)^{-1}, \tag{54}$$

which is the same expression as the RHS of (49).

6.2 Stratified Sampling

In this subsection the computation of the matrix $T(\hat{V}/P)T$ is given for a special case: (1) the disturbances are normally distributed, and (2) the sampling design is an endogenously stratified sampling design, such that the inclusion probability π_{ii} of element i of the population is a function f of only the i -th element of y , say $y_{(i)}$. Thus,

$$\pi_{ii} = f(y_{(i)}), \tag{55}$$

for $i = 1, \dots, N$. As an example, consider the stratified sample which was shown in Figure 1. The design contains three strata there. The elements in the middle stratum have the highest inclusion probability. Figure 2 shows the corresponding function f .

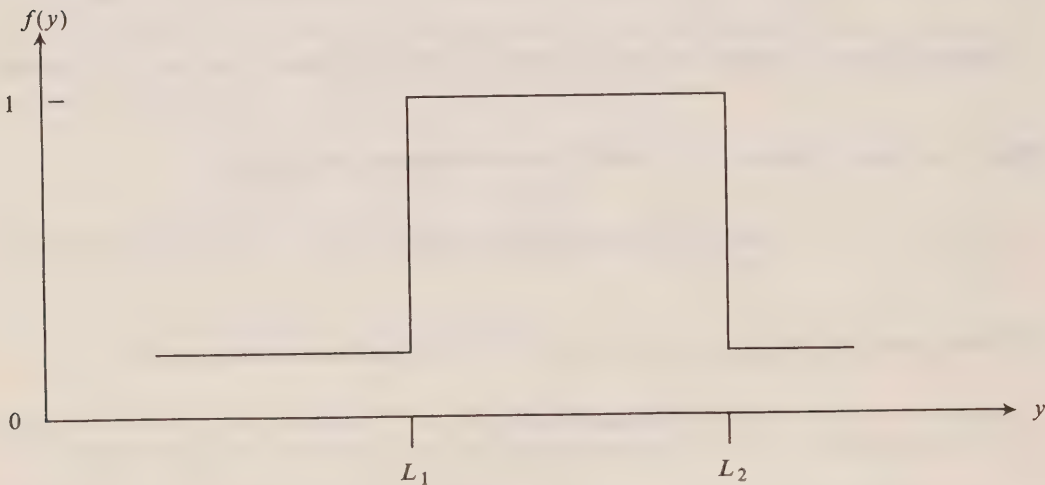


Figure 2. The Probability Function f Corresponding to Figure 1

In general, let there be H strata, indicated by $h = 1, \dots, H$. Let the boundaries of these strata be L_0, L_1, \dots, L_H . Typically, $L_0 = -\infty$ and $L_H = +\infty$. Let $\pi_{(h)}$ be the inclusion probability of the population elements in stratum h . More formally, the function $f(\cdot)$ is such that $f(y)$ equals $\pi_{(h)}$ if $L_{h-1} \leq y < L_h$. The values of $\pi_{(h)}$ and L_h are usually known in practice, since the actual sampling design depends on their values.

In stratified sampling, the second order inclusion probability of any two population elements not in the same stratum equals the product of their respective first order inclusion probabilities: their inclusions in the sample are independent. For any two population elements in the same stratum this holds approximately. Thus, approximately the off-diagonal elements of P are equal to the off-diagonal elements of $\Pi\iota'\Pi$. The diagonal of P is equal to the diagonal of Π , as before. Thus, approximately,

$$P = \Pi\iota'\Pi - \Pi^2 + \Pi. \quad (56)$$

Then

$$\begin{aligned} V &= E_{\xi} [\text{diag}(\epsilon)(\iota\iota' - I + \Pi^{-1}) \text{diag}(\epsilon)] \\ &= E_{\xi} [\epsilon\epsilon' - \text{diag}^2(\epsilon) + \text{diag}^2(\epsilon)\Pi^{-1}] = E_{\xi} [\text{diag}^2(\epsilon)\Pi^{-1}], \end{aligned} \quad (57)$$

in view of assumption (4). Thus V is a diagonal matrix here. Then

$$T\left(\frac{V}{P}\right)T = T\Pi^{-1}E_{\xi}[\text{diag}^2(\epsilon)\Pi^{-1}], \quad (58)$$

which is also a diagonal matrix. Now consider a population element i , which is included in the sample. Then, using (58) and assuming normally distributed disturbances,

$$\begin{aligned} \left[T\left(\frac{\hat{V}}{P}\right)T\right]_{ii} &= \frac{1}{\pi_{ii}} \sum_{h=1}^H \frac{1}{\pi_{(h)}} \int_{L_{h-1}-x_i'\hat{\beta}}^{L_h-x_i'\hat{\beta}} \varphi(\epsilon_i; \hat{\sigma}^2) \epsilon_i^2 d\epsilon_i \\ &= \frac{\hat{\sigma}^2}{\pi_{ii}} \left\{ \frac{1}{\pi_{(H)}} + \sum_{h=1}^{H-1} \left(\frac{1}{\pi_{(h)}} - \frac{1}{\pi_{(h+1)}} \right) \Psi[(L_h - x_i'\hat{\beta})/\hat{\sigma}] \right\}. \end{aligned} \quad (59)$$

Here, $\phi(\cdot; \hat{\sigma}^2)$ indicates the normal density with mean zero and variance $\hat{\sigma}^2$. The function $\Psi(\cdot)$ is defined as

$$\Psi(x) \equiv \int_{-\infty}^x \varphi(\epsilon; 1) \epsilon^2 d\epsilon = \Phi(x) - x\varphi(x; 1), \quad (60)$$

where $\Phi(\cdot)$ denotes the cumulative density function for the standard normal distribution. In the derivation of (59), use has been made of $\Psi(L_0) = 0$ and $\Psi(L_H) = 1$.

7. MODEL-FREE REGRESSION

7.1 Consistent Estimation

As a digression from the main theme of this paper, model-free regression will be considered in this section. Firstly, model-free regression can be usefully applied in the case of doubt about the validity of a linear model. See Fuller (1975), who studies model-free regression for some specific designs. Van Praag (1981, 1982) studies model-free regression in the

case of repeated sampling from some probability distribution. See also DuMouchel and Duncan (1983). White (1980b, Section 3) studies related problems. Secondly, the so-called regression estimator of a population total uses model-free regression. See textbooks such as Cochran (1977), the review papers mentioned above by Nathan (1981) and Smith (1981) and Bethlehem and Keller (1983).

The purpose of model-free regression is the estimation of the population parameter vector

$$b \equiv (X'X)^{-1}X'y, \quad (61)$$

without assumptions about the probability distribution of y . In fact, both X and y are considered non-stochastic. Further, the same replica structure as in Section 2 is used, as follows.

$$X = \iota_K \otimes X_0, \quad (62)$$

and

$$y = \iota_K \otimes y_0, \quad (63)$$

where y_0 is some fixed N_0 -vector. As before, the K diagonal matrices T_k ($k = 1, \dots, K$) are i.i.d. These matrices describe the sample as in Section 2. Together the matrices T_k form the matrix T . No additional assumptions are made concerning the distribution of T .

It is proved relatively easily, along the same lines as in Section 2, that the weighted estimator $\hat{\beta}$ defined before in (7), is a consistent estimator of b defined in (61). See also Jönrup and Rennermalm (1976), who indicates $\hat{\beta}$ as an "approximately unbiased" estimator of b , and Van Praag (1982, Section 4d), where "selectivity bias" with known inclusion probabilities is studied for the model-free case.

It follows in the same manner as in Section 4 that in the model-free case the asymptotic variance of $\hat{\beta}$, say $\text{Var}_{\text{MF}}(\hat{\beta})$, equals

$$\text{Var}_{\text{MF}}(\hat{\beta}) = (X'X)^{-1}X'VX(X'X)^{-1}, \quad (64)$$

with

$$e \equiv y - Xb, \quad (65)$$

$$V = \text{diag}(e)\Pi^{-1}P\Pi^{-1}\text{diag}(e), \quad (66)$$

and with P defined as before in (29). Notice that V in (66) differs from V in (28) in the omission of the ξ -expectation and the substitution of e for ε .

It is interesting to rewrite $\text{Var}_{\text{MF}}(\hat{\beta})$ in the same way as $\text{Var}(\hat{\beta})$ was rewritten in Section 5. In doing so, use will be made of

$$X'e = 0, \quad (67)$$

which follows directly from (61) and (65). The $\text{Var}_{\text{MF}}(\hat{\beta})$ can be rewritten as

$$\begin{aligned} \text{Var}_{\text{MF}}(\hat{\beta}) &= (X'X)^{-1}X'\text{diag}(e)(\Pi^{-1}P\Pi^{-1} - u'u')\text{diag}(e)X(X'X)^{-1} \\ &\quad + (X'X)^{-1}X'ee'X(X'X)^{-1} \\ &= (X'X)^{-1}X'\text{diag}(e)(\Pi^{-1}P\Pi^{-1} - u'u')\text{diag}(e)X(X'X)^{-1}. \end{aligned} \quad (68)$$

The last member of (68) corresponds with the p -component of the decomposition of $\text{Var}(\hat{\beta})$ in (37). It may be concluded from (68) that in model-free regression the variance of the estimator of the regression coefficients consists of the p -component, while the ξ -component vanishes.

Notice finally that, using the discussion at the end of Section 5, the last member of (68) can be written as

$$(X'X)^{-1}\Sigma(X'X)^{-1}, \quad (69)$$

where the matrix Σ is the p -variance-covariance of the row totals of $X'\text{diag}(e)$. A similar result was reached by Binder (1983, Section 4), though along different lines.

8. DISCUSSION

In this section some practical considerations are given concerning the use of weights in regression analysis. Several motives for the use of weights are discussed shortly, related to the preceding technical sections of this paper.

First of all, it must be noted that the difference between weighted and unweighted regressions may be of some significance. An important example is the case where business firms are the unit of study – either farms, industrial enterprises of any other kind of business firms varying considerably in the number of employees. At the Netherlands Central Bureau of Statistics, for instance, the classification by number of employees is a standard stratification variable in sampling designs of business firms, giving a considerable range of inclusion probabilities – the large units chosen with relatively large probabilities. In studies with employment as the endogenous variable, such a sampling design is endogenous, which calls for weighted regression; the large units receiving small weights.

Secondly, in the case of units varying widely in size, a major problem with regression analysis is the heteroscedasticity of the error term. This calls for weighted regression, of the same sort as the weighting due to an endogenous design discussed in Section 2: large units receiving small weights.

Finally, there is a third motive for the weighting of sampled data: the notion of a model free regression, as discussed in Section 7 above. Again, the weights here are of the same sort as the weights in Section 2.

Summing up, there seems to be no reason not to incorporate the sampling design in regression analysis.

9. CONCLUSIONS

In this paper the estimation of a regression model with survey sample data has been studied. In particular, samples drawn with an endogenous design have been studied; for example, a sample stratified on the endogenous variable. It has been shown that for such a sample the weighting of the observations with the inverse of the square root of the sampling fractions gives a consistent estimator. The concept of consistency used here is a modification of Brewer (1979). The asymptotic variance of the estimator has been given, as well as a consistent estimator of this variance. The variance is the sum of a sampling component and a model component.

Also, model-free regression has been considered. Model-free regression requires the same weighting as endogenous stratification. The variance of the estimator of the model-free regression coefficients contains only the sampling component, and not the model component.

Finally, some practical considerations relative to the weighting of the data have been given.

ACKNOWLEDGEMENT

The author thanks Abby Israëls and Albert Verbeek and several anonymous referees for their comments on previous versions of this paper.

REFERENCES

- BETHLEHEM, J.G., and KELLER, W.J. (1983). Weighting sample survey data using linear models. Internal Report, Department for Statistical Methods, Netherlands Central Bureau of Statistics, Voorburg.
- BINDER, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BRESLOW, N., and DAY, N.E. (1980). *Statistical Methods in Cancer Research, Volume 1: the Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.
- BREWER, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.
- CRAMER, J.S. (1971). *Empirical Econometrics*. Amsterdam: North-Holland.
- DuMOUCHEL, W.H., and DUNCAN, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, C37, 117-132.
- HAUSMAN, J.A., and WISE, D.A. (1981). Stratification on endogenous variables and estimation: the Gary income maintenance experiment. In *Structural Analysis of Discrete Data with Econometric Applications*, (Eds., C.F. Manski and D. McFadden), Cambridge: MIT Press.
- HECKMAN, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-161.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- JEWELL, N.P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, 72, 11-21.
- JOHNSTON, J. (1972). *Econometric Methods*. Tokyo: McGraw-Hill Kogakusha.
- JONRUP, H., and RENNERMALM, B. (1976). Regression analysis in samples from finite populations. *Scandinavian Journal of Statistics*, 33-36.
- KMENTA, J. (1978). *Elements of Econometrics*. New York: McMillan.
- MANSKI, C.F., and MCFADDEN, D. (eds.) (1981). *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: MIT Press.
- NATHAN, G. (1981). Notes on inference based on data from complex sample designs. *Survey Methodology*, 7, 110-129.
- PORTER, R.D. (1973). On the use of survey sample weights in the linear model. *Annals of Economic and Social Measurement*, 2, 141-158.
- RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley.

- ROBINSON, P.M. (1982). On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics*, 24, 234-238.
- SMITH, T.M.F. (1981). Regression analysis for complex surveys. In *Current Topics in Survey Sampling*, (Eds. D. Krewski, R. Platek, and J.N.K. Rao), New York: Academic Press, 267-292.
- THEIL, H. (1971). *Principles of Econometrics*. New York: Wiley.
- VAN PRAAG, B.M.S. (1981). Model-free regression. *Economics Letters*, 7, 139-144.
- VAN PRAAG, B.M.S. (1982). The population-sample decomposition with an application to minimum distance estimators. Report 8218, Center for Research in Public Economics, Leyden University.
- WHITE, H., (1980a). Nonlinear regression on cross section data. *Econometrica*, 48, 721-746.
- WHITE, H., (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, 12, 149-170.

A Cluster Analysis of Activities of Daily Living From the Canadian Health and Disability Survey¹

D.A. BINDER and G. LAZARUS²

ABSTRACT

The Canadian Health and Disability Survey, administered as a supplement to the Canadian Labour Force Survey in October 1983, collected data on potentially disabled persons by means of a screening questionnaire and a follow-up questionnaire for those screened-in. The data from the screening questionnaire, consisting of a set of activities of daily living, were used to group respondents according to identifiable characteristics. A description of the groups of respondents is provided along with an evaluation of the methods used in their determination. An incompletely ordered severity scale is proposed.

KEY WORDS: Disability scale; Discriminant analysis.

1. INTRODUCTION

Considerable efforts have been made to acquire a better understanding of the disabled population. These efforts have focussed on the development of a useful vehicle for capturing the potentially disabled population as well as the analysis of survey data for the purposes of gaining a better understanding of the various dimensions of disability and to develop useful measures of severity. Examples of papers which examine these issues are Dolson *et al.* (1984) and Raymond *et al.* (1981), among others. This paper chronicles the development of an exploratory technique in order to gain a better understanding of the disabled population in Canada. In particular, a cluster analysis based on results of several discriminant analyses was performed.

The next section presents information about the Canadian Health and Disability Survey. The third section describes the development of the clusters. Section 4 focusses on the characterization of the clusters. Some analysis of the behaviour of the derived clusters is given in Section 5. The paper concludes with some closing remarks.

2. BACKGROUND

In response to a need for data on disabled persons in Canada, Statistics Canada undertook a program to create a disability database. The Canadian Health and Disability Surveys (CHDS) were administered as supplements to the Canadian Labour Force Survey (LFS) in October 1983 and June 1984. In both cases, separate questionnaires were administered to children and to adults. In the October survey, the adult questionnaire was administered to everyone in the LFS sample (the frame includes about 97% of the Canadian population aged 15 or more). In June, the adult survey was restricted to those aged 15 to 64 from the six provinces with the smaller sample sizes in October (i.e. Newfoundland, Prince Edward Island, Nova Scotia, New Brunswick, Manitoba and Saskatchewan). Children from all provinces were surveyed in both October and June.

¹ This is a revised version of the paper presented at ASA meetings, Social Statistics Section, Las Vegas, August 1985.
² D.A. Binder and G. Lazarus, Social Survey Methods Division, Informatics and Methodology Branch, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

This paper concentrates on work which utilized only the data from the adults questionnaire in October 1983. This survey obtained 92,945 adult respondents from approximately 47,000 households.

2.1 Questionnaire

2.1.1 Screening Section

The Labour Force Supplement included a screen which was used to identify respondents for a follow-up questionnaire. The screening section consisted of nineteen items – seventeen activities of daily living, an activity limitation item and an item about mental handicap. The activities of daily living (ADL's) are a set of activities which any person would perform during the course of his/her regular living pattern. The set used here was a modified version of those developed by the Organization for Economic and Co-operative Development (OECD) and has been utilized by several other countries.

The ADL's are presented in Table 1 with the questionnaire identification and the orientation of the specified activity. Two ADL's are related to hearing troubles, two to vision troubles, four to mobility troubles, one to speaking and being understood and the remaining eight to agility troubles.

Table 1
Activities of Daily Living

Questionnaire Item	Description	Orientation
A10	Walking 400 Metres	Mobility
A11	Walking up and down stairs	Mobility
A12	Carrying 5 kg. object for 10 metres	Mobility
A13	Moving from one room to another	Agility
A14	Standing for long periods	Mobility
A15	When standing, bending down to pick up object	Agility
A16	Dressing and undressing	Agility
A17	Getting in and out of bed	Agility
A18	Cutting own toenails	Agility
A19	Using fingers to grasp or handle	Agility
A20	Reaching	Agility
A21	Cutting own food	Agility
A22	Reading newsprint	Vision
A23	Seeing clearly a face across the room	Vision
A24	Hearing conversation with another person	Hearing
A25	Hearing conversation with two or more persons	Hearing
A26	Speaking and being understood	Speaking and being understood

An example of the wording of these questions in the screening section of the questionnaire is as follows: (A20) Does . . . have any trouble reaching? The activity limitation item (A27) concerned limitation "in the kind or amount of activity he/she can do at home, at work or going to school because of a long-term physical condition or health problem". The final item in the screen section (A28) concerned mental handicap.

It should be noted that the survey was concerned with long-term conditions or health problems – those that had lasted or were expected to last more than six months (excluding pregnancy). An individual was screened in if he/she had trouble with at least one of the ADL's, the activity limitation item or had a mental handicap. (Proxy responses were required for mentally handicapped individuals).

2.1.2 Follow-up Section

The follow-up section of the questionnaire was completed for individuals selected by the screening section. This section included an item which sought to determine if the respondent was completely unable to perform the ADL('s) he/she had trouble with. Other segments of the follow-up questionnaire pertained to: nature of the disability (related to trouble seeing or reading, trouble hearing, trouble speaking and being understood, and mobility); problems related to the ability to work or the workplace itself; obstacles to education and availability of special educational facilities; problems related to local and long-distance travel; and problems in current residence and special facilities. The information in the follow-up questionnaire, given above, could be used to analyze the cluster characteristics, or to develop a severity index (see Lazarus; 1985a, 1985b).

3. CLUSTERS

This section presents a description of the procedures used in the development of the clusters. The clustering procedures employed were developed specifically for this application. Technical details concerning the methods used are given in Sections 3.2 and 3.3. All computations were performed using SAS.

3.1 Methodology

This section summarizes the methodology used to derive the final clusters. The clustering procedure consisted of two steps:

- a) a divisive step, where the 12,907 individuals were sequentially partitioned using PROC CANDISC.
- b) an agglomerative step, where the partition was collapsed.

For the divisive step, the following procedure was employed iteratively. First, the starting point put all the observations into a single cluster. Each step subdivided each of the current clusters into two groups. For each of the current clusters, a canonical correlation analysis was performed by taking each non-constant variable as a grouping variable and using all other non-constant variables as explanatory variables. The cluster was then split into two, based on the discriminant analysis with the largest F -value. In this way the determinant of the between-sums-of-squares matrix is maximized.

For the agglomerative step, subjective criteria were used, based on the magnitude of the F -value, the size of the groups and the plots of the points. Collapsing was accomplished in the reverse order of splitting, for the most part.

For the divisive step, data based on both unweighted and weighted covariances were used separately. The results were essentially the same. It was decided to continue without the sampling weights because of the added complexity which would be incurred by their inclusion. Furthermore, the weights were not expected to be important with respect to the characteristics of the clustered individuals. Inclusion of weights is necessary for evaluation and analysis.

3.2 Description

The cluster analysis was a procedure which grouped together those screened in respondents with similar but not necessarily identical “profiles”. For our purposes, a respondent’s profile consisted of the responses to the seventeen ADL’s (yes, has trouble/no, does not have trouble), responses to the major activity limitation item (positive/negative), and the mental handicap item in the screening section of the questionnaire.

Table 2 details the final clusters. The symbols *U* and *Z* demonstrate how the groups are defined. The symbol *U* means that the group is defined through that variable being one, i.e. 100% by definition. The symbol *Z* is used when the defining screening section item is zero, i.e. 0% by definition. Note that six of the nineteen screening items are not used explicitly in the process of classifying respondents. These are A11, A13, A18, A20, A23 and A24.

4. CLUSTER CHARACTERIZATION

This section explores the ways and means of identifying the clusters. The concepts of “trouble orientation” and “umbrella” group are introduced and the clusters are ranked according to the severity of disability.

4.1 Trouble Orientation

Threshold values were established to assist in the cluster classification process. The values were chosen by ordering the clusters according to orientation and locating an obvious gap in the *E(NADL)* for the orientation, where *E(NADL)* referred to the average number of troubles among ADL’s A10 - A26. In general, a cluster was recognized as having trouble with an activity orientation when the *E(NADL)* for a particular orientation exceeded the established threshold value. For example, for mobility orientation, *E(NADL)* was computed for activities A10, A11, A12 and A14. The *E(NADL)* for each cluster over each orientation may be found in Table 3.

Clusters were labelled as follows. If a cluster had trouble with an activity, the corresponding letter was included in the label. Two clusters, containing individuals who had trouble speaking and being understood or were mentally handicapped, were “special”. Clusters which had neither mobility nor agility troubles exceeding the established values were so designated with an *N*. For example, HMA1 and HMA2 refer to clusters with a large proportion having hearing, mobility and agility problems, but no particular problem with vision. Alternatively, VN1 refers to a cluster with the exact opposite set of problems.

4.2 Umbrella Groups

Clusters with similar orientation patterns became members of specified “umbrella” groups, where they could be better compared using *E(NADL)* within the umbrella. Table 4 shows the clusters according to the “umbrella” groups to which they belong.

Table 2
Cluster Analysis Results

Cluster	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
1	U	92.7	79.9	59.7	89.8	85.5	U	62.7	86.8	60.1
2	U	77.0	63.1	16.0	77.0	55.6	Z	11.2	46.5	31.0
3	U	85.1	66.5	19.4	75.8	U	Z	15.8	49.6	26.5
4	U	65.6	36.7	6.4	55.9	Z	Z	4.5	21.5	17.7
5	Z	18.7	18.2	3.4	25.6	24.6	4.9	6.9	21.7	20.7
6	Z	36.3	23.2	4.8	49.5	U	11.8	16.6	28.4	21.1
7	Z	9.2	5.3	0.3	10.8	Z	1.1	0.9	4.4	7.1
8	U	94.7	88.6	67.3	93.9	89.0	U	74.7	94.7	84.0
9	U	92.9	82.1	55.4	89.3	91.1	U	58.9	87.5	30.4
10	U	95.7	81.0	55.7	91.9	93.8	U	U	85.2	33.3
11	U	92.2	71.7	21.1	83.7	74.1	U	Z	59.0	28.9
12	U	91.9	71.3	25.0	81.3	U	Z	16.9	58.1	31.9
13	U	61.0	48.8	4.3	55.5	Z	Z	4.9	32.3	14.0
14	U	91.3	U	23.6	81.4	U	Z	16.8	40.2	U
15	U	93.6	U	29.9	84.0	U	Z	19.3	56.1	Z
16	U	74.9	Z	10.9	65.7	U	Z	12.9	32.8	16.4
17	U	66.7	58.3	12.5	37.5	Z	Z	0.0	37.5	20.8
18	U	74.0	55.5	7.5	59.5	Z	Z	10.4	29.5	U
19	U	79.6	U	11.5	60.8	Z	Z	2.9	14.6	Z
20	U	59.0	Z	2.7	45.6	Z	Z	2.2	10.4	Z
21	Z	14.7	12.6	1.9	19.4	13.9	5.5	4.7	22.2	11.5
22	Z	26.5	40.9	7.0	41.4	59.1	U	32.1	47.4	35.8
23	Z	29.0	26.1	2.1	43.3	U	Z	13.0	19.0	13.5
24	Z	2.4	2.4	0.0	2.0	Z	Z	0.4	7.7	3.3
25	Z	35.6	U	2.4	32.9	Z	Z	3.1	8.5	18.0
26	Z	13.5	Z	0.3	16.8	Z	Z	1.8	4.2	9.1
27	Z	17.0	13.7	0.3	U	Z	Z	2.4	6.2	5.4
28	Z	10.3	6.9	0.0	Z	Z	Z	0.1	7.8	U
29	Z	38.7	26.3	0.6	Z	Z	Z	2.2	10.9	Z
Cluster	A20	A21	A22	A23	A24	A25	A26	A27	A28	Size
1	63.7	42.2	38.6	27.1	73.3	U	23.4	94.4	6.3	303
2	35.3	11.8	U	50.8	71.7	U	9.6	85.0	1.6	187
3	34.6	5.9	Z	3.4	63.7	U	2.5	88.7	1.1	355
4	16.4	1.9	Z	1.6	57.9	U	2.6	73.3	1.0	311
5	17.7	8.4	U	46.3	59.6	U	12.8	55.7	7.9	203
6	24.9	3.5	Z	1.4	50.9	U	4.2	71.3	1.0	289
7	4.6	0.6	Z	1.3	60.5	U	5.6	26.3	1.6	1,770
8	78.4	U	32.6	16.7	1.2	Z	32.2	96.3	9.8	245
9	50.0	Z	U	30.4	5.4	Z	10.7	100.0	5.4	56
10	55.2	Z	Z	0.5	0.0	Z	2.4	89.0	1.9	210
11	45.8	Z	Z	1.8	0.6	Z	3.0	90.4	0.6	166
12	39.4	7.5	U	45.6	4.4	Z	5.0	93.1	1.9	160
13	20.7	5.5	U	42.7	1.2	Z	6.7	78.0	4.3	164
14	34.4	1.5	Z	1.0	0.9	Z	1.3	89.4	1.2	187
15	66.3	16.6	Z	2.1	2.1	Z	5.9	92.0	1.6	677
16	20.7	0.7	Z	0.0	0.4	Z	2.0	82.3	0.4	458
17	16.7	20.8	Z	0.0	0.0	Z	U	91.7	33.3	24
18	29.5	12.1	Z	0.0	0.0	Z	Z	82.1	1.2	173
19	19.4	1.0	Z	0.5	0.2	Z	Z	73.5	1.0	582
20	8.0	0.0	Z	0.7	0.4	Z	Z	66.7	0.6	857
21	9.7	7.1	U	41.1	2.6	Z	8.7	55.3	9.2	618
22	41.9	19.5	Z	1.4	1.4	Z	7.0	76.3	4.7	215
23	18.1	1.9	Z	0.8	0.7	Z	1.2	66.6	0.4	1,164
24	0.8	2.0	Z	0.0	0.0	Z	27.2	62.2	U	246
25	23.7	1.4	Z	0.3	0.0	Z	1.4	U	Z	295
26	7.3	1.2	Z	0.7	0.5	Z	1.9	U	Z	1,923
27	2.4	0.3	Z	0.3	0.3	Z	0.3	Z	Z	371
28	11.8	8.3	Z	0.0	0.5	Z	0.5	Z	Z	204
29	18.0	1.6	Z	6.5	5.7	Z	8.5	Z	Z	494

Table 3
Average Number of Troubles by Orientation

Cluster	Hearing	Vision	Mobility	Agility	Total
1	1.733	0.657	3.624	5.841	11.855
2	1.717	1.508	3.171	2.170	8.566
3	1.637	0.034	3.274	2.543	7.488
4	1.579	0.016	2.582	0.710	4.887
5	1.596	1.463	0.625	1.211	4.895
6	1.509	0.014	1.091	2.152	4.766
7	1.605	0.013	0.253	0.246	2.117
8	0.012	0.493	3.772	7.203	11.480
9	0.054	1.304	3.643	4.480	9.841
10	0.000	0.005	3.686	5.256	8.947
11	0.006	0.018	3.476	3.319	6.819
12	0.044	1.456	3.445	2.838	7.783
13	0.012	1.427	2.653	0.884	4.976
14	0.009	0.010	3.727	3.178	6.924
15	0.021	0.021	3.776	2.941	6.759
16	0.004	0.000	2.406	1.964	4.374
17	0.000	0.000	2.625	2.083	4.708
18	0.000	0.000	2.890	1.890	4.780
19	0.002	0.005	3.404	0.494	3.905
20	0.004	0.007	2.046	0.233	2.290
21	0.026	1.411	0.467	0.852	2.756
22	0.014	0.014	1.088	3.498	4.614
23	0.007	0.008	0.984	1.688	2.687
24	0.000	0.000	0.068	0.352	0.482
25	0.000	0.003	1.685	0.587	2.273
26	0.005	0.007	0.303	0.258	0.573
27	0.003	0.003	0.310	1.170	1.486
28	0.005	0.000	0.172	1.285	1.462
29	0.057	0.065	0.650	0.418	1.190

4.3 Severity

One area of analytic interest is the development of an index of severity of disability. The notion has been considered previously by Raymond et al, among others.

The index of severity would be useful in as much as it would allow for simple comparisons of disability among the screened-in respondents. The use of *E(NADL)* to draw such comparisons presumes that the orientations are self-weighting, noting, for example, that two ADL's are devoted to hearing troubles while four are devoted to mobility troubles. Also, the multidimensional nature of severity of disability is hidden by a single score such as *E(NADL)*.

Table 4
Ordering of Clusters by "Umbrella" Groups

Umbrella Group	Cluster	Sample Count	<i>E</i> (NADL)	ID
HV (Hearing/Vision)	2	187	8.566	HVMA1
	5	203	4.895	HVN1
H (Hearing)	1	303	11.855	HMA1
	3	355	7.488	HMA2
	4	311	4.829	HM1
	6	289	4.760	HA1
	7	1,770	2.120	HN1
V (Vision)	9	56	9.841	VMA1
	12	160	7.783	VMA2
	13	164	4.976	VM1
	21	618	2.756	VN1
S (Special)	17	24	4.708	SMA1
	24	246	0.482	SN1
MA (Mobility/Agility)	8	245	11.480	MA1
	10	210	8.947	MA2
	11	166	6.819	MA4
	14	187	6.924	MA3
	15	677	6.759	MA5
M (Mobility)	16	458	4.374	M2
	18	173	4.780	M1
	19	582	3.905	M3
	20	857	2.290	M4
A (Agility)	22	215	4.614	A1
N (Neither)	23	1,164	2.687	N1
	25	295	2.273	N2
	26	1,923	0.573	N6
	27	371	1.486	N3
	28	204	1.462	N4
	29	494	1.190	N5

Table 4 presents an ordering of clusters according to "severity" within umbrella groups. This within group ordering better reflects the notion that severity is multidimensional than would an overall ordering.

5. CLUSTER CHARACTERISTICS

The principal components technique was used to examine the behaviour of the resulting clusters. Raymond et al also employed principal components; the main difference being that analysis here is based upon group means rather than individuals.

5.1 Methodology

We considered a subset of screened in cases, where more information per case is available. In particular, we added the responses to questions of the form: (B101) Is . . . completely

unable to walk 400 metres without resting? This line of questioning was used for each of the ADL'S, A10-A26. Thus, 11,412 of the original 12,907 individuals who were screened in were usable. The other 1,495 were dropped because of non-response problems. These "completely unable" items were coded with "1" when the individual indicated that he/she was completely unable to perform the specified ADL, otherwise, a "0" was coded.

The means were obtained for the nineteen screening items and seventeen follow-up items for each cluster. The means for the completely unable items were then multiplied by the ratio of the overall average number of ADL's to the overall average of completely unable items in order to scale them consistently and to avoid the scaling problems associated with principal components analysis.

Principal components were obtained using the nineteen screening section and seventeen follow-up item means as variables, using the "clusters" as observations and weighting according to cluster size. The clusters were then ordered according to each of the first four principal component scores.

The final stage involved the pooling of cluster cases according to "umbrella" group membership and finding the means of the first four principal component leadings for each of the eight "umbrella" groups, where the weights were the numbers of members in the "umbrella" groups.

5.2 Results

We present the results in two stages. In the first stage, we examine the principal components and attempt to label them according to the scores. We also explore the "umbrella" group construct in terms of the principal component means. In the second stage, we examine the ordering of the clusters according to the first four principal components.

5.2.1 Components

The first four principal components for the nineteen screening section items and the seventeen follow-up items explained just over seven-eighths of the total variance and appeared to be most useful for our purposes.

The loadings of the first principal component are positive on all but four items (A24, A25 and B241 are hearing oriented, A28 is mental handicap). The negative loadings are close to zero. This first component appears to be an overall measure of strength. The first principal component explained nearly 66% of the total variance and is denoted as "OVERALL".

There are negative loadings on A10, A11, A12, A14 and A15 of the second component. The loading for A15 is nearly zero, however. Loadings are positive for ADL's with an agility-trouble orientation as well as for hearing-trouble and vision-trouble orientations. It appears then that this component polarizes mobility trouble against agility, hearing and vision troubles. The second component is labelled "AHV/M".

The third principal component has positive loadings for mobility and hearing oriented ADL's and negative loadings for agility and vision oriented ADL's. This third component is denoted "MH/AV".

The fourth principal component has positive loadings for mobility and vision oriented ADL's and negative loadings for agility oriented ADL's. This fourth component is designated "MV/A".

5.2.2 Mean Loadings

Table 5 presents the average differences of the principal component scores from the mean scores over all 11,412 individuals, for each of the eight "umbrella" groups. We can

Table 5
Average Differences of Principal Component
Scores from Mean Scores

Umbrella Group	Sample Count	Differences			
		PRIN1 (Overall)	PRIN2 (AVH/M)	PRIN3 (MH/AV)	PRIN4 (MV/A)
Hearing/Vision	346	0.68	1.26	0.61	1.06
Hearing	2741	-0.33	0.54	0.81	-0.25
Vision	888	0.30	0.69	-0.76	1.27
Special	151	-1.02	-0.04	-0.47	-0.06
Mobility/Agility	1311	3.31	-0.33	-0.21	-0.33
Mobility	1893	0.30	-0.80	0.18	0.33
Agility	195	-0.19	0.31	-0.80	-0.78
Neither	3887	-1.11	-0.16	-0.41	-0.22

now check to see if the incomplete ordering presented earlier is consistent with the results from the principal components analysis. We note the following observations are taken from Table 5.

- i) The mobility/agility “umbrella” group has the highest difference on the first principal component “overall”, while the “umbrella” group “neither” has the lowest difference. The difference for the hearing/vision group is positive as is the mean for the vision group. The hearing group difference is negative, however, evidence that individuals with hearing-oriented troubles tend not to have other disabilities. There may be an inclination to draw the same kind of conclusion with respect to agility-oriented troubles. It is observed that the mobility/agility and mobility groups have positive differences while the agility “umbrella” group has a negative difference. However, in this case, the result is somewhat ambiguous because the agility-oriented ADL’s included speaking trouble (A26), a so-called “special” trouble area and it is clear indeed that the special “umbrella” group has a negative difference for the first principal component.
- ii) The second component set mobility-oriented troubles (-) against agility, hearing and vision-oriented troubles (+). Positive differences are recorded for the hearing/vision, hearing, vision and agility “umbrella” groups while negative differences are associated with the mobility/agility, mobility and neither groups, as expected. The difference for the special groups is nearly zero.
- iii) The third component set mobility-oriented and hearing-oriented troubles (+) against agility-oriented and vision-oriented troubles (-). Again, the results are consistent.
- iv) The fourth principal component set mobility and vision-oriented troubles (+) against agility-oriented troubles (-). The results are again consistent with the umbrella-group construct.

5.2.3 The Scales

Table 6 shows the ranks of the clusters according to the first four principal component scores and E(NADL). Recall that the component loadings are for 11,412 cases and utilize follow-up information as well as screening section information while the E(NADL) scale is based on 12,907 cases and uses screening information only.

The cluster ranking according to principal components was done as follows. The component representing overall strength (OVERALL) ranked clusters from highest to lowest scores. The ranking of clusters on AHV/M tended to put clusters with mobility-oriented troubles at the bottom end as opposed to clusters with agility, hearing or vision oriented troubles

which were ranked higher up on this scale. The ranking of clusters on MH/AV tended to put clusters with mobility or hearing troubles at or near the bottom of the scale while clusters with agility or vision-oriented troubles were ranked higher. Finally clusters with agility-oriented troubles were ranked higher on MV/A than the others. Given the bipolar nature of components 2, 3 and 4, it was necessary to make an arbitrary decision as to a trouble orientation scale. As cluster 8 had shown itself to be highly severe according to the *E(NADL)* scale, it was determined that cluster 8 should be similarly ranked along the other scales.

For most clusters, the rankings fluctuate over a wide range. This reflects the nature of the criteria upon which the scales were based. The first principal component, which provides an overall measure of strength, may be the most suitable candidate for ranking the clusters. Firstly, it incorporates the screening section information used in the development of the *E(NADL)* measure. As a result, the rank orderings provided by the OVERALL and *E(NADL)* scales are quite similar. The additional follow-up information used in the construction of

Table 6
Cluster Rank According to Alternative Scales

Cluster	ID	PRIN1 (Overall)	PRIN2 (AHV/M)	PRIN3 (MH/AV)	PRIN4 (MV/A)	<i>E(NADL)</i>
2	HVMA1	9	4	27	28	5
5	HVN1	22	2	22	25	12
1	HMA1	3	3	24	6	1
3	HMA2	10	14	28	10	7
4	HM1	16	15	29	20	13
6	HA1	20	8	25	3	15
7	HN1	29	7	26	9	24
9	VMA1	2	6	4	23	3
12	VMA2	4	10	7	27	6
13	VM1	13	11	11	29	11
21	VN1	23	5	2	26	20
8	MA1	1	1	1	1	2
10	MA2	5	20	13	4	4
14	MA3	6	24	16	7	8
11	MA4	7	23	17	8	9
15	MA5	8	28	20	18	10
18	M1	14	26	19	21	14
16	M2	15	25	18	17	18
19	M3	11	29	23	24	19
20	M4	18	27	21	22	22
22	A1	17	9	3	2	17
23	N1	21	17	6	5	21
25	N2	19	22	10	16	23
27	N3	24	19	15	12	25
28	N4	28	12	9	11	26
29	N5	25	16	12	15	27
26	N6	26	18	8	14	28
17	SMA1	12	21	14	19	16
24	SN1	27	13	5	13	29

this component leads us to believe that OVERALL is better than other scales such as *E(NADL)*. It is worth noting that the ranking was done on all 29 clusters and depicted in Table 6 on an "umbrella" group basis. The "umbrella" group information was not incorporated into the principal components analysis, however.

6 CLOSING REMARKS

A clustering technique was employed to group screened-in individuals according to similar screening section profiles. The clusters were then ordered according to the information contained in the screening section of the questionnaire (the incomplete ordering based on *E(NADL)* and presented in Table 4) and finally according to information contained in the screening and follow-up sections of the questionnaire (the OVERALL scale presented in Table 6). This last scale is deemed presently to be the most suitable of those considered here. However, it could be argued that no single index of severity exists and in fact the severity index should be defined as a 4-dimensional scale corresponding to our principal components.

ACKNOWLEDGEMENT

The authors would like to thank the assistant editors for constructive comments on earlier drafts.

REFERENCES

- DOLSON, D., GILES, P., and MORIN, J.-P. (1984). A methodology for surveying disabled persons using a supplement to the Labour Force Survey. *Survey Methodology*, 10, 187-197.
- LAZARUS, G. (1985a). Characteristics of potentially disabled individuals based on the cluster analysis of activities of daily living. Working Paper, Institutional and Agricultural Survey Methods Division, Statistics Canada.
- LAZARUS, G. (1985b). An application of the results of the cluster analysis of activities of daily living. Working Paper, Institutional and Agricultural Survey Methods Division, Statistics Canada.
- RAYMOND, L., CHRISTE, E., and CLEMENCE, A. (1981). Vers l'établissement d'un score global d'incapacité fonctionnelle sur la base des questions de l'OCDE, d'après une enquête en Suisse. *Revue d'épidémiologie et santé publique*, 29, 451-459.

Additive Versus Multiplicative Seasonal Adjustment When There Are Fast Changes in the Trend-Cycle¹

GUY HUOT and NAZIRA GAIT²

ABSTRACT

The seasonal adjustment of a time series is not a straightforward procedure particularly when the level of a series nearly doubles in just one year. The 1981-82 recession had a very sudden great impact not only on the structure of the series but on the estimation of the trend- cycle and seasonal components at the end of the series. Serious seasonal adjustment problems can occur. For instance: the selection of the wrong decomposition model may produce underadjustment in the seasonally high months and overadjustment in the seasonally low months. The wrong decomposition model may also signal a false turning point. This article analyses these two aspects of the interplay between a severe recession and seasonal adjustment.

KEY WORDS: Decomposition models; ARIMA; Lead-lag relationship.

1. INTRODUCTION

1981 and 1982 were atypical years afflicted by a severe recession. This recession has profoundly affected the evolution and structure of economic time series, and consequently their seasonal adjustment. Seasonally adjusted time series are necessary to diagnose the socio-economic health of a country. In turn, social and economic policies founded on these data influence decisions in both the private and public sectors. Thus, this recession raises many questions. One can readily see that a prompt examination of seasonal adjustment is necessary.

The series under consideration here are: initial and renewal claims received (for unemployment benefits) and beneficiaries. It is difficult to see how their trend and cycle components evolve when they are contaminated by seasonal variation, namely intra-annual climatic and institutional factors. Seasonal adjustment permits a better detection of fundamental tendencies, such as turning points, and evaluation of the present performance of the economy.

This article analyses some aspects of the interplay between a severe recession and seasonal adjustment. In just one year, that is in 1981, this recession has nearly doubled the level of beneficiaries. Such a sudden large change prompts questions about the structure of the series, the choice of the X-11-ARIMA decomposition model, the determination of turning points at the end of the series, and the use of ARIMA forecasts for seasonal adjustment.

In section 2, we discuss two important consequences of using a wrong decomposition model, namely a systematic over- and under-adjustment of series and the possibility of having a false turning point at the end of the series. In section 3, we use the lead-lag relationship between the claims and beneficiaries series to help seasonally adjust the latter series.

The ARIMA forecasts generally help to reduce the revision to the seasonal factors and they can help to provide a more accurate recognition of the turning points at the end of the series. Section 4 considers this question.

¹ This paper was presented at the 145th Annual Meeting of the American Statistical Association, Las Vegas, Nevada, 1985.

² Guy Huot, Time Series Research and Analysis Division, Statistics Canada. N. Gait, University of Sao Paulo, Brazil, was visiting Statistics Canada when the paper was written.

2. DECOMPOSITION MODELS FOR SEASONAL ADJUSTMENT

Most of the claims and beneficiaries series have similar characteristics, so we have chosen to study one claims series and one beneficiaries series which can clearly illustrate some of the problems peculiar to seasonal adjustment during a severe recession. It should be noted that the results of our analysis are equally valid during a sudden strong expansion in the economy. It is the sudden large change in the level of the series caused by the recession or the expansion that is important.

The X-11-ARIMA program (Dagum 1980) will be used to seasonally adjust these series. The program is applied to the claims and beneficiaries series, using data from January 1973 and May 1975 respectively, up to February 1983.

The X-11-ARIMA program provides three decomposition models for the estimation of the time series components. The program assumes an additive relationship between the components

$$O_t = TC_t + S_t + I_t \quad (2.1)$$

or a multiplicative one

$$O_t = TC_t S_t I_t \quad (2.2)$$

or a log additive one

$$\log O_t = \log TC_t + \log S_t + \log I_t \quad (2.3)$$

where O stands for the observed and unadjusted series; TC , the trend-cycle; S and I , the seasonal and irregular components; and t , the time.

Seasonal adjustment means removing the seasonal variations S_t from the raw data O_t , thus leaving a seasonally adjusted series consisting of TC_t and I_t . In order to know whether a certain series contains a significant amount of seasonality and if so, whether an additive or multiplicative model provides the better fit, one can perform a test for the presence of seasonality and a model test on the series (Higginson 1977). The first test shows that both series contain a very significant amount of seasonality. According to the second test, the multiplicative model fits the beneficiaries series better when tested from May 1975 to June 1981. When the series is extended to February 1983, taking into account the impact of the recession on the series, the additive model then fits better. On the other hand, the model test favours neither the additive nor the multiplicative model for the claims series.

One usually adjusts the series using only one model, however, figure 1 shows the beneficiaries series adjusted using the two models, both without using the ARIMA option. During 1980 and 1981, the difference between the additive and multiplicative adjustments was small compared with the difference observed in 1982.

The multiplicative model assumes that the seasonal variation is proportional to the level of the trend-cycle. During 1982, the seasonal amplitude did not increase in this way. Consequently, using the multiplicative model is likely to overestimate it from June to November, the seasonally low months. As figure 1 shows, in underestimating the number of seasonal beneficiaries, the multiplicative model has drastically overestimated the number of seasonally adjusted beneficiaries. The converse is also true.

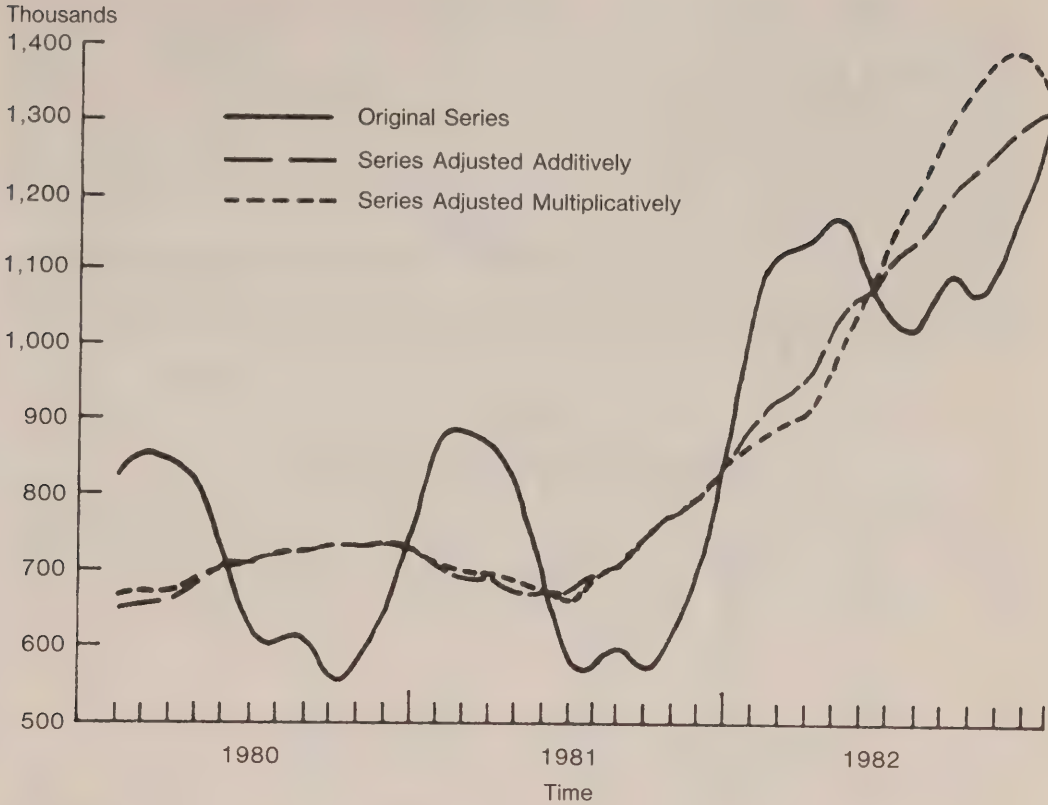


Figure 1. Beneficiaries

The additive model, on the other hand, does not assume that the components of the series evolve proportionately. Figure 1 confirms that the trend cycle increased while the seasonal amplitude remained constant. Thus, the additive model provides the better seasonal adjustment. It performs better in 1982 than the multiplicative model and is acceptable in 1980 and 1981.

By mid-1982, it was not easy to tell which of the additive or multiplicative models would adjust the beneficiaries series better. Since this series was adjusted multiplicatively until June 1981, one would normally continue to do so in 1982. During 1982, were there some clues or pieces of evidence showing that the multiplicative model was no longer adequate?

The acceptance or rejection of model, given a sudden large change in the level of a series, clearly has to be based on a thorough analysis of the data. The set of quality control statistics included in the X-11-ARIMA program is not meant to detect that kind of problem in the model. In this experiment with the multiplicative model, none of the ten individual control statistics failed the guideline. However, the F test for the presence of moving seasonality showed the presence of increasing moving seasonality during 1982 in the final unmodified SI ratios.

Besides a systematic over and under-adjustment of the series, another consequence of using a wrong decomposition model is the possibility of having a false turning point at the end of the series.

Let us say that a cyclical turning point has occurred if the seasonally adjusted series shows a change in direction that persists for at least 5 months. Once the beneficiaries series has been seasonally adjusted multiplicatively, figure 1 shows the possible presence of a turning point around October 1982, where the upward trend has suddenly changed to a downward trend. This turning point seems to be confirmed when the series ending in December 1982 is extended by one month. The additively adjusted series, on the other hand, shows no turning point. The two results conflict. Thus, either the multiplicative model is signaling a false turn or the additive model is missing the turning point.

It is not that easy to show that the multiplicative model has signalled a false turn. The multiplicative model has created a turning point around October 1982. Table 1 shows that in the very short run, the updating of the series did not reverse this turning point.

Table 1
Multiplicatively Adjusted Beneficiaries Series
(in thousands, July 1982 – February 1983)

July	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.
124	131	140					
124	130	140	142				
124	130	140	141	141			
124	130	140	142	138	131		
123	130	140	142	141	131	121	
123	129	139	142	141	134	121	123

3. LEAD-LAG RELATIONSHIP BETWEEN THE CLAIMS AND BENEFICIARIES SERIES

Leading indicators are sensitive to the evolution of the economic climate. They are measures of anticipations or new commitments, and as such they give an advance indication of changes expected in the trend-cycle of coincident and lagging indicators.

Figure 2 shows the claims series as a leading indicator for the beneficiaries series. The performance of the seasonally adjusted indicators can be tested using the criteria of Klein and Moore (1982). The two series satisfy these criteria. First, the correspondence between the series is one-to-one – the number of cycles is the same in each series. Second, there is uniformity in timing – the claims series always lead. Third, these are monthly series and they are current, or up-to-date. Thus, the claims series is likely to predict an upward or a downward change in the trend of the beneficiaries series.

The lead-lag relationship between the two series can help to seasonally adjust the beneficiaries series. It reduces the likelihood of mistaking an irregular turn for a cyclical turning point. Figure 2 shows September 1982 to be a turning point in the multiplicatively adjusted claims series. This is also true for the additive adjustment of the series. Since the cross-correlations between the two series shows a lead-lag relationship of 5 to 6 months, the September 1982 turning point in the claims series indicates that the multiplicative model applied to the beneficiaries series has signalled a false turn around October 1982. However, the leading indicator predicts a turning point around March 1983 in the beneficiaries series.

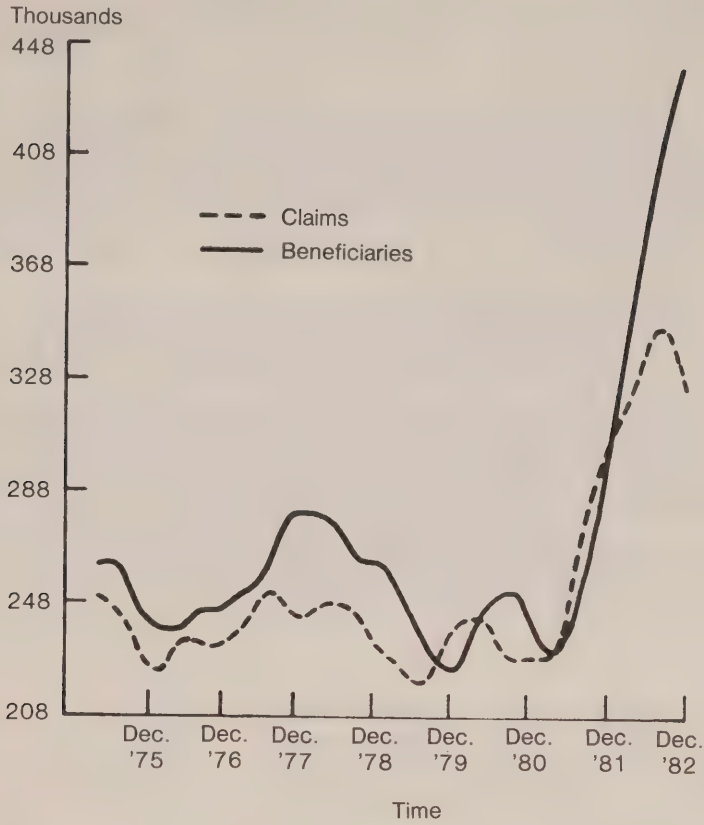


Figure 2. Claims and Beneficiaries. The Number of Beneficiaries has been Divided by 3 in Order to Make the Scale of Both Series Compatible.

4. ARIMA EXTRAPOLATIONS

An optimal seasonal adjustment procedure has to minimize the revision to the current seasonal factors and also has to produce reliable estimates of the trend-cycle, particularly of turning points, at the end of the series (Dagum 1979). The analysis carried on in the previous sections is based on seasonally adjusted data without using the ARIMA option. In this section, we shall focus on the use of the ARIMA forecasts as a variable that can provide an accurate recognition of the turning points.

The automatic X-11-ARIMA program proceeds as follows:

- 1. Three univariate ARIMA models of the general multiplicative form $(p,d,q) (P,D,Q)_s$ (Box and Jenkins 1970) are fitted to the monthly or quarterly series that is to be seasonally adjusted. The models are

$$\begin{aligned} &(0,1,1) (0,1,1)_s \\ &(0,2,2) (0,1,1)_s \\ &(2,1,2) (0,1,1)_s \end{aligned}$$

when the series is seasonally adjusted additively. For series adjusted multiplicatively, the same models are used and the log transform is applied to the data for the first two models.

2. The series is extrapolated one year in advance; and
3. provided the extrapolations are acceptable, the ordinary X-11 method is then applied to the series thus extended.

Figure 3 shows the beneficiaries series seasonally adjusted both additively and multiplicatively, using the automatic X-11-ARIMA options. The ARIMA models that best fit and forecast the series ending in December 1982 are (0,2,2) (0,1,1)₁₂ when the series is seasonally adjusted additively and log (0,2,2) (0,1,1)₁₂ when adjusted multiplicatively. The log (0,2,2) (0,1,1)₁₂ model has forecast a decrease in the series, while the (0,2,2) (0,1,1)₁₂ model has maintained the upward trend.

Figure 3. shows the multiplicative seasonal adjustment of the beneficiaries series using both the upward trend and the downward trend extrapolations. One can see from the comparison of figure 1 with figure 3 that ARIMA extrapolations did not modify the multiplicative estimates of the trend-cycle in the last year. The multiplicative model is still signalling a turning point around October (downward trend, log transform). The multiplicative model applied to either the non-extended beneficiaries series (figure 1) or to the extended series is questionable.

By the end of 1983, one could see that the true turning point has actually occurred around February 1983. Thus, the October or November 1982 turning point can hardly be corrected by extrapolation when it is due to the wrong selection of the decomposition model.

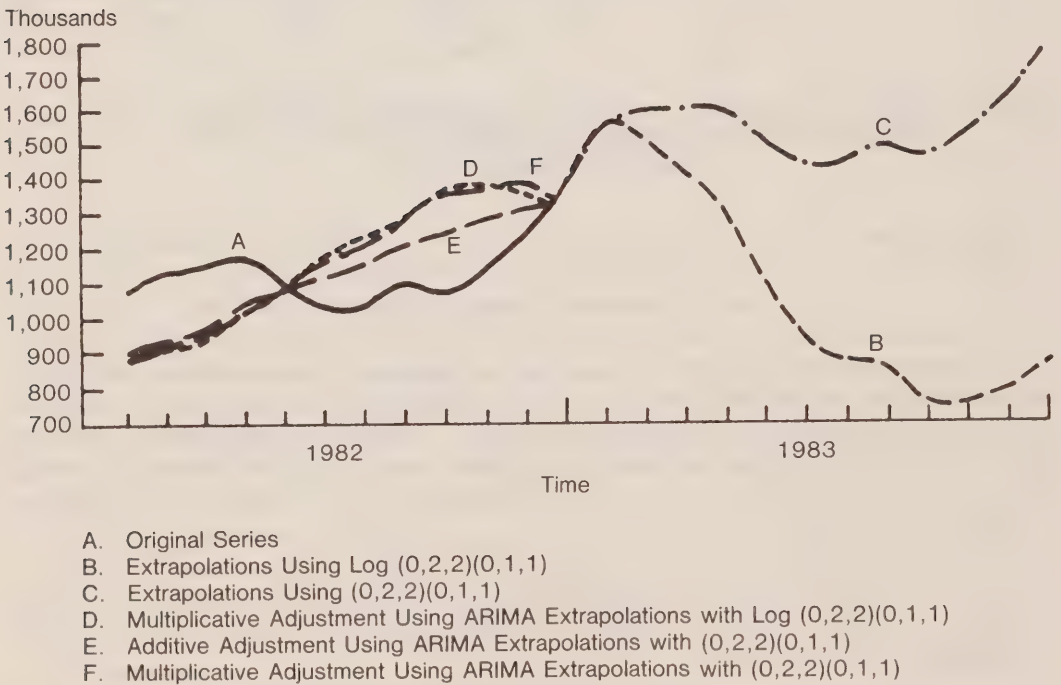


Figure 3. Beneficiaries Series Seasonally Adjusted Additively and Multiplicatively with Different ARIMA Extrapolations

Over and under-adjustment and problems of identifying the turning points occurred in other series as well. Figure 4 shows for instance, the series of "benefits paid" when seasonally adjusted multiplicatively with actual data available to the end of 1984. The seasonally adjusted series tends to oscillate systematically around the trend-cycle curve at the turning point, thus over- and underestimating the benefits paid. After the turning point, the oscillation decays to the trend-cycle curve; showing that the multiplicative model is doing poorly around the turning point. Note that this series has strong trading-day-variation which has also been removed.

5. SELECTION OF THE OPTIMAL SEASONAL ADJUSTMENT PROCEDURE

Figure 5 summarizes the criteria for seasonal adjustment that have been taken into account to overcome the problems due to the interplay between the 1981-82 recession and seasonal adjustment of the beneficiaries and claims series. The selection of the best seasonal adjustment procedure was primarily based on the first criterion.

In order to avoid over- and underestimation and false turning points in the seasonally adjusted figures, the appropriate decomposition model has to be selected. A thorough analysis of the data should be conducted by:

1. performing a model test on the series.
2. adjusting the series both additively and multiplicatively if the effort is justified. If the difference between the two adjustments becomes significant as in figure 1, one has to check for underadjustment in the seasonally high months and for overadjustment in the seasonally low months. One can also look in table D8 of the X-11-ARIMA program at the F tests for the presence of stable and moving seasonality. The decomposition model that better adjusts the series will usually show the higher F value for stable seasonality and the lower F value for moving seasonality.
3. checking for turning points. For the claims series, both decomposition models have signalled a turn in August or September 1982. On the other hand, for the beneficiaries series, only the multiplicative model has signalled a turn in October 1982. Thus either the multiplicative model is signalling a false turn or the additive model is missing the turning point. The analysis has shown this turn to be a false one resulting from the drastic overestimation of the number of seasonally adjusted beneficiaries in the seasonally low months as shown in Figure 1.
4. using a bi- or multivariate approach to accurately estimate the turning points at the end of the series. The lead-lag relationship between the claims and beneficiaries series can help to seasonally adjust the beneficiaries series. It reduces the likelihood of mistaking an irregular turn for a cyclical turning point. Since the lead is about 5 to 6 months, the September 1982 turning point in the claims series confirms that the multiplicative model applied to the beneficiaries series has signalled a false turn in October 1982. However, the leading indicator is predicting a turning point around March 1983 in the beneficiaries series.
5. using the ARIMA option with concurrent seasonal factors. It usually gives smaller revisions to the seasonal factors whether an additive or a multiplicative seasonal adjustment is made. However, a false turning point can hardly be corrected by extrapolations when it is due to the wrong selection of the decomposition model.

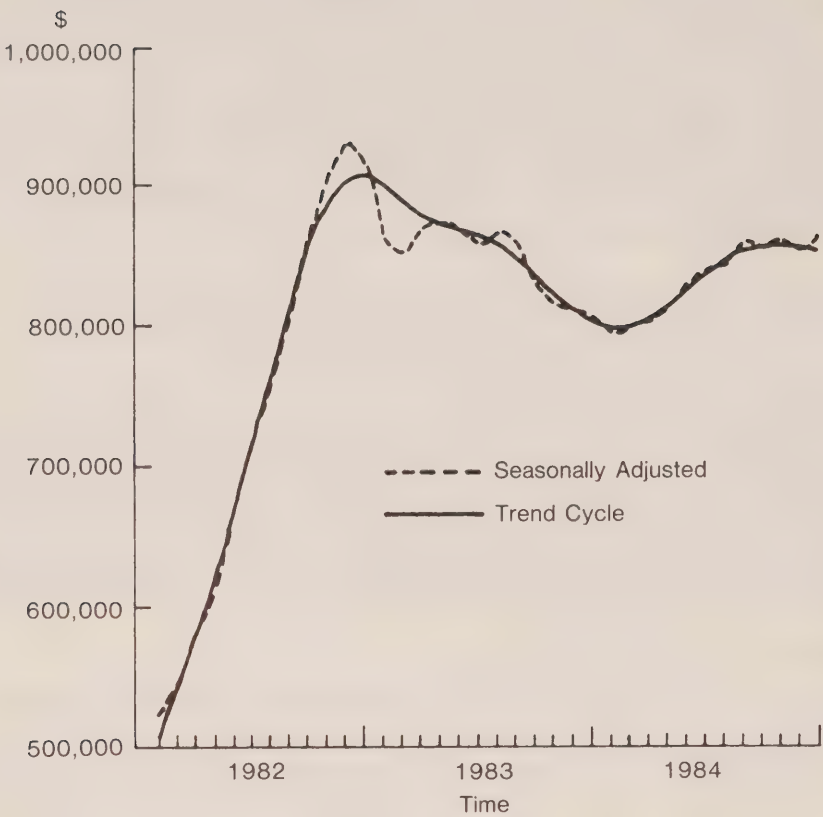


Figure 4. Benefit Paid (Seasonally Adjusted Multiplicatively)

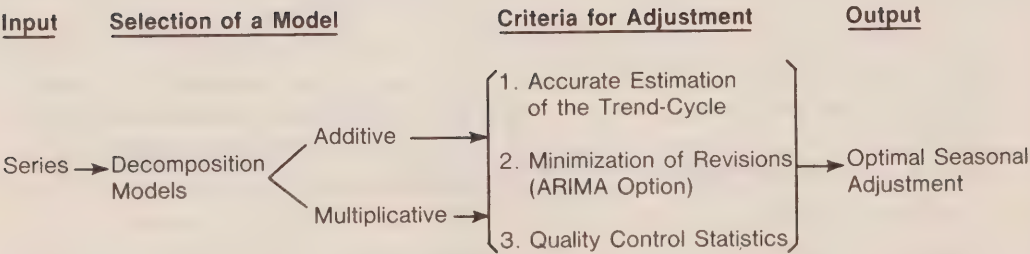


Figure 5. Optimal Seasonal Adjustment Procedure

6. checking both the raw and seasonally adjusted data. One cannot rely on tests only. For instance, the set of quality control statistics included in the X-11-ARIMA program is not meant to detect under- or overestimation of the series or false turning points.
7. all the above recommendations apply if the series is not strongly affected by trading-day-variation. If trading-day-variation is present, then it must be removed before the ARIMA option is used.

REFERENCES

- BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.
- DAGUM, E.B. (1979). Data extrapolation and smoothing with the X-11-ARIMA seasonal adjustment method. *Proceedings of the 12th Annual Symposium Interface Computer Science and Statistics*, (ed. Jane F. Gentleman), University of Waterloo, 195-202.
- DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method*. Statistics Canada, Catalogue No. 12-564E.
- HIGGINSON, J. (1977). *User Manual for the Decomposition Test*. Time Series Research and Analysis Division, Statistics Canada, Reference No. 77-01-001.
- KLEIN, P.A., and MOORE, G.H. (1982). *The Leading Indicator Approach to Economic Forecasting Retrospect and Prospect*. Center for International Business Cycle Research, Rutgers University, Newark, N.J.

Nonresponse Adjustment Procedures at the U.S. Bureau of the Census

DAVID W. CHAPMAN, LEROY BAILEY, and DANIEL KASPRZYK¹

ABSTRACT

Nearly all surveys and censuses are subject to two types of nonresponse: unit (total) and item (partial). Several methods of compensating for nonresponse have been developed in an attempt to reduce the bias associated with nonresponse. This paper summarizes the nonresponse adjustment procedures used at the U.S. Census Bureau, focusing on unit nonresponse. Some discussion of current and future research in this area is also included.

KEYWORDS: Nonresponse adjustments; Imputation; Missing data; Weighting.

1. INTRODUCTION

The Bureau of the Census has long recognized the potential seriousness of measurement errors ascribed to survey nonresponse, and has consistently incorporated nonresponse adjustment or compensation procedures in the estimation methodologies for its numerous and varied surveys and censuses. The objectives of this paper are to provide an overview of procedures employed by the Census Bureau in compensating for nonresponse, primarily unit nonresponse. By unit nonresponse we mean that little or no information for the principal survey variables is obtained for the sample unit in question.

This presentation will include (1) a discussion of the general weighting scheme used for the demographic surveys; (2) a review of some of the distinct problems associated with nonresponse in the Survey of Income and Program Participation (SIPP); (3) a discussion of the handling of unit nonresponse for the economic surveys and censuses; and (4) a section on imputation for earnings for the Current Population Survey. In addition to providing descriptions of the various nonresponse compensation methods used by the Census Bureau, the authors will cite specific problems associated with those methods and note the Bureau's current nonresponse research activities and concerns.

2. NONRESPONSE IN DEMOGRAPHIC SAMPLE SURVEYS

At any given time, the Bureau of the Census may be involved with the conduct of 25-30 recurring or special demographic surveys. The concerns of these surveys include labor force participation, individual and family income, health care, transportation, leisure activities, crime, and other topics reflective of the current interests of the nation's people, governments, businesses, and institutions. Unit nonresponse rates for these surveys range from between three and four percent for the National Crime Survey to over 25 percent, which was recorded for the 1984 National Survey of Natural and Social Scientists and Engineers.

¹ David W. Chapman and Leroy Bailey are Principal Researchers, Statistical Research Division, U.S. Bureau of the Census, Washington D.C. 20233. Daniel Kasprzyk is a Special Assistant, Office of the Chief, Population Division, U.S. Bureau of the Census, Washington D.C. 20233.

Weight adjustment within classes (Oh and Scheuren 1983), or cell balancing, is the predominant technique used to compensate for unit nonresponse in the Census Bureau's demographic surveys. There is variation among the surveys relative to the determination of adjustment classes within which weighting occurs. For some surveys, ancillary data available to define weighting classes are limited to basic geographic and survey design information, while for others a considerable amount of demographic and economic data are accessible.

The nonresponse adjustment factors for the Bureau's demographic surveys are usually the inverse of the survey's weighted or unweighted response rate. In a small number of the surveys this adjustment factor is modified slightly to reflect information gleaned from follow-up subsamples of the initial nonrespondents. Since the Census Bureau's general approach to survey nonresponse is essentially the same for all of its major demographic surveys, a general description will be given in Section 2.1 of the nonresponse adjustment procedure for the National Crime Survey (NCS), as the example of a "typical" Census Bureau application of weighting. Section 2.2 will consist of a discussion of alternative procedures and current unit nonresponse research in the demographic areas.

2.1 The National Crime Survey

The NCS sample is a national probability sample of about 72,000 households which is divided into six panels, each of which is interviewed in a given month and again at six-month intervals over three years. The survey focuses on measuring household crimes and the extent of victimization of household members age 12 and older by assault (including rape), burglary, larceny, auto theft, and robbery. [For a detailed description of the NCS, see U.S. Department of Commerce, Bureau of the Census (1977).]

Estimates for the NCS, which are produced quarterly, are derived by initially inflating the sample data by the inverse of the related selection probabilities. The noncontacts and refusals account for about three to four percent of the survey's occupied units in any given month. Adjustments for these units are made by applying adjustment factors to the weighted respondent data in weighting classes. An attempt is made to define these classes in such a way that the respondents and nonrespondents in each class have similar survey characteristics. In order to temper the impact of the nonresponse adjustment on the variance of the survey estimates, some of the smaller weighting classes generally have to be collapsed with other classes before a final nonresponse adjustment can be effected. Collapsing of classes also takes place if the weight adjustment factor becomes too large for one or more classes. [See Hanson (1978).] Collapsing is discussed further in Section 4.

Since the NCS employs a self-response method of interviewing, there is concern about the amount of within household nonresponse. Consequently, a separate set of weighting cells exists to compensate for within-household nonresponse. These cells or weighting classes, as well as those used for the household nonresponse adjustment, are indicated in Tables 1-3. The NCS household and within household nonresponse rates for 1984 are shown in Table 4.

To illustrate the NCS estimator of a total, there is a selection probability $\pi_i = 1, 2, \dots, N$, associated with each of the N units in the population. It is assumed that among the n sample units, n_R are respondents. The NCS estimator for the population total, after adjusting for unit nonresponse, takes the following form:

$$\hat{Y}_{\text{NCS}} = \sum_{j=1}^M \sum_{k=1}^P (z_j u_k)^{-1} \sum_{\ell=1}^{n_{Rjk}} \frac{y_{j k \ell}}{\pi_{j k \ell}},$$

Table 1
NCS Noninterview Adjustment Cells for
Within Household Nonresponse

Household Relationship	Persons by Age, by Race of Head							
	Black				Non-black			
	12-24	25-44	45-64	65 +	12-24	25-44	45-64	65 +
Head of Household								
Wife of Head								
All other Persons								

Table 2
NCS Household Noninterview
Adjustment Cells for
Standard Metropolitan
Statistical Areas (SMSA's)

Race	Central City of SMSA	Balance of SMSA	
		Urban	Rural
White			
Not White			

Table 3
NCS
Household Noninterview
Adjustment Cells for
Non-SMSA's

Race	Urban	Rural	
		Non-farm	Farm
White			
Not White			

where for sample units in the k^{th} within household and j^{th} household weighting classes,

- y_{jkl} = value of the ℓ th sample respondent,
- n_{Rjk} = number of sample respondents,
- n_{jk} = number of sample cases,
- z_j = the estimated household response rate,
- u_k = the estimated within household response rate,
- π_{jkl} = selection probability for the ℓ th sample respondent,
- P = total number of within household nonresponse weighting classes,
- M = total number of household nonresponse weighting classes.

Implicit in the formation of the NCS nonresponse weighting classes, as well as those for other demographic surveys, are the following assumptions:

1. There is "significant" correlation between the major survey variables and the covariates used to define noninterview adjacent classes.
2. Within each household nonresponse weighting class, $E(\bar{y}_{Rj}) = E(\bar{y}_{Rj})$, where \bar{y}_{Rj} and \bar{y}_{Rj} are the means for the sample respondents and nonrespondents, respectively, in the j^{th} weighting class.
3. The weighting class means differ, that is, $E(\bar{y}_{Rj}) \neq E(\bar{y}_{Rj}'), j \neq j'$.

(Assumptions analogous to 2 and 3 above are also implicit for *within* household nonresponse adjustment classes.)

Table 4
NCS Noninterview Rates - 1984

	Average 1984	Jan.	Feb.	Mar.	Apr.	May	June
Household Noninterviews							
Total Interviewed HH's	11,769	11,916	11,925	11,743	11,809	11,918	9,482
Total	430	446	540	481	446	388	348
Rate	3.5	3.6	4.3	3.9	3.6	3.2	3.5
No one at home	0.9	0.8	1.1	0.9	0.9	0.7	1.0
Temporarily Absent	0.6	0.6	0.6	0.8	0.6	0.4	0.7
Refusal	1.9	2.1	2.6	2.2	2.2	2.0	1.9
Other	0.1	0.2	0.2	0.1	0.1	0.1	0.1
Within Household Noninterviews							
Total	685	655	751	701	806	804	697
Rate	2.5	2.6	3.0	2.8	3.0	2.9	3.2
		July	Aug.	Sept.	Oct.	Nov.	Dec.
Household Noninterviews							
Total Interviewed HH's		9,869	9,446	9,895	9,350	9,692	9,410
Total		411	409	337	406	387	346
Rate		4.0	4.2	3.3	4.2	3.8	3.5
No one at home		0.9	0.9	0.6	1.0	1.2	1.0
Temporarily Absent		1.0	1.0	0.6	0.6	0.4	0.4
Refusal		2.1	2.3	2.0	2.4	2.1	2.1
Other		0.1	0.1	0.1	0.3	0.3	0.1
Within Household Noninterviews							
Total		709	678	666	728	735	803
Rate		3.1	3.1	2.9	3.4	3.3	3.7

The selection of weighting classes for this procedure is constrained by the requirement that measurements for the weighting class variables (covariates) must be available (either before or during the survey) for both the respondents and the nonrespondents. This essentially restricts the characteristics by which classes are defined to those associated with geography, race, urbanicity, housing unit characteristics, and design levels. The bias reduction capability of the procedure depends, in part, on the extent to which the NCS nonresponse weighting classes satisfy the three assumptions given above. No definitive results relating to this concern are currently available, but relevant research is underway and more empirical studies seem warranted.

2.2 Alternatives to Sample Weighting

There are a number of plausible alternatives to weighting to adjust for nonresponse. See, for example, Little (1986, Section 5). However, there are no definitive results which show that any of them offer appreciable advantages. Subsections 2.2.1 and 2.2.2 contain brief descriptions of two alternatives which are currently being investigated for application to demographic surveys.

2.2.1 Separate Estimates for Dissimilar Types of Nonresponse

In demographic surveys, nonrespondents can be placed into four categories: refusal (REF), not-at-home (NAH), other occupied unit (OTO), or a unit from which a response was not obtained due to extenuating circumstances. These are referred to as type A noninterviews. The NAH group can be divided into those households or individuals whose extended absence from their homes precludes an interview during the scheduled interview period (NAH_E), and the group which is expected to return home sometime during the survey period (NAH_S).

The authors are not aware of any data which show that the four nonresponse groups are generally similar. In fact, the Census Bureau's Current Population Survey and the Canadian Labour Force Survey suggest that the NAH_S households are likely to be smaller, younger, and have a larger proportion of employed people than the other groups. The NAH_E group is usually older with a relatively low employment rate. The interviewed group may be more reflective of the REF and OTO groups. [See Palmer and Jones (1967) and Paul and Lawes (1982).] It is conceivable that separate treatment of the four nonresponse groups could produce a better overall adjustment for nonresponse than is obtained from the current procedure. This option is being investigated by an NCS nonresponse adjustment research group.

2.2.2 Weighting With Response Probabilities

Several weighting techniques have been advanced which make use of the concept of response probabilities. Most of these techniques are based on concepts introduced by Politz and Simmons (1949) which group sample respondents according to estimates of their probabilities of responding. The factors with which the sample data in the resultant weighting groups are inflated are the inverses of the estimated response probabilities. The Politz-Simmons procedure has some serious limitations, such as its inapplicability to refusals. However, there have been a number of fairly recent extensions and applications of the procedure, including those presented by Anderson (1978), Thomsen and Sirling (1983). These methods may be applicable to recurring surveys for which extensive callbacks are made.

Research is in progress regarding the development of models which may be used to estimate response probabilities for several demographic surveys for units with similar values of the "independent variables." The feasibility and merits of computing nonresponse adjustment factors, as well as constructing weighting classes based on such models (sometimes referred to as response propensity stratification), are being examined. [See Rosenbaum and Rubin (1983) and Little and Samuhel (1983).] Moreover there are continued efforts to develop more objective methods of sample weighting for nonresponse, which are designed to control nonresponse-related errors.

3. THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

The Survey of Income and Program Participation (SIPP) is a new, ongoing national household survey program of the U.S. Bureau of the Census. The purpose of SIPP is to improve the measurement of information related to the economic situation of households

and persons in the United States. It is the culmination of a large-scale development program, the Income Survey Development Program (ISDP), which examined concepts, procedures, questionnaires, recall periods, and the like. For a description of the ISDP, see Ycas and Linger (1981). Data from SIPP are expected to be useful in studying the Federal transfer system, estimating program costs under changes in program eligibility rules, evaluating the effects of program changes on selected population subgroups, as well as studying changes to the tax system.

In October 1983 SIPP began as an ongoing survey program with one sample panel of approximately 21,000 occupied households eligible for interview in 174 Primary Sample Units (PSU's) selected to represent the noninstitutional population of the United States. (Beginning in 1985 a new panel is being introduced in February of each year; the 1985 panel consisted of 14,500 households eligible for interview.)

Each household is interviewed once every four months for approximately 2½ years to produce sufficient data for longitudinal analysis while providing a relatively short recall period for reporting monthly income. The reference period for the principal survey items is the 4 months preceding the interview. This design provides eight interviews per household, and allows cross-sectional estimates to be produced from more than one panel.

To facilitate field and processing operations, each sample panel is divided into four approximately equal subsamples, called rotation groups; one rotation group is interviewed in a given month. Thus, one cycle or "wave" of interviewing, using the same questionnaire, takes four consecutive months. Cumulative *household* noninterview rates are given in Table 5 for the 1984 SIPP panel.

At the time of the interviewer's visit, each person 15 years old or older who is present is asked to provide information about himself/herself; a proxy respondent is asked to provide information for those who are not available. An important design feature of SIPP is that all persons in a sample household at the time of the first interview remain in the sample even if they move to a new address during the next 2½ years. For cost and operational reasons, in-person interviews are only conducted at new addresses that are within 100 miles of a SIPP primary sampling unit. The geographic areas defined by these rules contain over 96% of the U.S. population. An attempt is made to conduct a telephone interview with those moving outside the 100-mile limit.

Table 5
Cumulative Household Noninterview Rates
for the 1984 SIPP Panels

Wave	Sample Loss
1	4.9%
2	9.4%
3	12.3%
4	15.4%
5	17.4%
6	19.4%
7	21.0%
8	22.0%
9	22.3%

After the first interview, the SIPP sample is a person-based sample, consisting of all individuals who were living in the sample unit at the time of the first interview. Individuals aged 15 and over who subsequently share living quarters with original sample people are also interviewed in order to provide the overall economic context of the original sample persons.

More detailed information concerning the SIPP design, content, and operations can be found in Nelson, McMillen, and Kasprzyk (1985).

3.1 Nonresponse Adjustments in SIPP

Data collected in SIPP can be viewed from two perspectives: cross-sectional or longitudinal. From the former point of view, each SIPP interview is treated as a separate cross-sectional survey, providing point-in-time estimates. For examples of these estimates, see U.S. Department of Commerce, Bureau of the Census (1984a). From the longitudinal point of view, data are collected at more than one point-in-time, and the survey record is viewed not as a set of unrelated observations, but as a set of variables with logical dependency between two or more points-in-time. Data processing operations, as well as statistical estimation, are treated from this point of view, and therefore, rely on the use of data collected at two or more interviews.

Since SIPP can be viewed from both the longitudinal and cross-sectional perspectives, SIPP's public-use microdata files include cross-sectional data files issued on a wave-by-wave basis as well as longitudinal files. This implies two distinct systems to treat survey nonresponse.

3.1.1 Cross-Sectional Unit Nonresponse Adjustments

The cross-sectional unit nonresponse adjustment in SIPP is similar to the way noninterview adjustments are made in other Census Bureau recurring surveys. The following variables were used to define household noninterview adjustment cells for the first interview wave of SIPP. See U.S. Department of Commerce, Bureau of the Census (1983 and 1984b).

1. Census Region – Northeast, Midwest, South, West.
2. Residence – Standard Metropolitan Statistical Area (SMSA), non-SMSA.
3. Place/not place – defined for units not in an SMSA,
Central city/balance – defined for units in SMSA's.
4. Race of reference person – black, non-black
5. Tenure – owner of home, renter.
6. Household size – 1, 2, 3, 4 or more.
7. Rotation group – 1, 2, 3, 4.

Two criteria must be met by each weighting class: (1) the weighting class must contain at least 30 unweighted units and (2) the noninterview adjustment factor for a weighting class must be less than or equal to 2.0. For a given rotation group, the collapsing procedure to satisfy these two criteria is applied independently for each of the four tenure by race combinations. (For the first wave, there was no *within*-household nonresponse adjustment factor.)

In subsequent waves of SIPP, the household nonresponse adjustment factor accounts for noninterviews associated with units which have moved and cannot be located or have moved more than 100 miles from a SIPP PSU and cannot be contacted by telephone as well as units which are refusals, etc. Adjustments are performed for each month of the reference period, as well as the interview month, to account for an increase in the number of noninterviews

caused by splits of sample households. The procedure is similar to that described for determining the Wave 1 household nonresponse adjustment factor; however, the variables used to define the weighting classes differ. Those variables are:

1. Race (white, nonwhite) and Spanish-origin (Spanish, non-Spanish) of reference person: a) reference person is white and not Spanish, and b) others.
2. Household type – three categories: a) female householder, no husband present, with own children under 16, b) householder's age is sixty-five years or older, and c) others.
3. Education level of the reference person: a) less than 8 years, b) 8-11 years, c) 12-15 years, and d) 16 or more years.
4. Type of income received (using the most recently completed interview for members of the household) – two categories: a) households which received at least one of the following sources of income – Supplemental Security Income; Black Lung Payments; Aid to Families with Dependent Children; General Assistance, Indian, Cuban, or Refugee Assistance; foster child care payment; Women's, Infants', and Children's Nutrition program; Food Stamps; and Medicaid; and b) others.
5. Assets – two categories: a) households in which at least one member held an asset type other than a savings account or an interest-bearing checking account, and b) all others.
6. Tenure: a) owner of home and b) renter.
7. Public housing or rent subsidies--renters are identified as a) those living in public housing projects or receiving rent subsidies from the government; and b) those not living in public housing projects and not receiving rent subsidies from the government.
8. Household size: 1, 2, 3, 4 or more.

The variables used for household nonresponse adjustments for the second and subsequent SIPP interviews differ from the first wave variables because of additional data available after the first interview for use in nonresponse procedures for later interviews. Fifty-three weighting classes were created using these variables with tenure as the principal variable for partitioning the sample. [For a description of these weighting classes see U.S. Department of Commerce, Bureau of the Census (1984c).] Although a cell collapsing strategy has been defined which merges cases in cells exhibiting similar poverty-related characteristics, little collapsing takes place since the nonresponse adjustment factors are calculated for three rotation groups (the SIPP data processing cycle) rather than one rotation group, as in the first interview.

There is a within-household nonresponse compensation procedure for the second and subsequent waves. This procedure is to "hot deck" (i.e., duplicate) the entire record of a sample respondent who presumably has survey characteristics that are similar to those of the nonrespondent.

3.1.2 Longitudinal Nonresponse Adjustments

Since persons identified as living at the sample address at the time of the first interview constitute the SIPP sample for waves subsequent to the first, the most useful and logical way of describing the nature of the SIPP nonresponse problem from the longitudinal viewpoint is in terms of individuals or persons. Each individual's microdata record is an extended record containing variables which oftentimes reflect the same measure at different points in time. Thus, in a panel survey of n waves there exist 2^n possible noninterview patterns for a sample person. Noninterview patterns of the original sample persons for the first five interviews (waves) of the 1984 panel are given in Table 6, adapted from Kalton, McMillen, and Kasprzyk (1986).

Table 6
Interview patterns of the Original Sample Persons for the First Five Interviews
of the 1984 SIPP Panel

Response Pattern	Percent
Response every interview (5 interviews)	
Pattern: XXXXX	79.1
Apparent attrition cases	13.8
Patterns: XXXXO	3.8
XXXOO	3.1
XXOOO	3.2
XOOOO	3.7
First and fifth interviews conducted, but one and more interven-	
ing interview missing	4.1
Patterns: XXXOX	1.6
XOXXX	0.6
XXOXX	1.2
XXOOX	0.1
XOXOX	0.1
XOOOX	0.3
XOOXX	0.2
Fifth interview missing and one or more intervening interviews	
missing	0.7
Patterns: XOXXO, XOXOO, XOOXO, XXOXO	
Left the universe (deceased, institutionalized, living in armed	
forces barracks, moved overseas)	2.3
Total	100.0
	(25,128)

The first SIPP longitudinal microdata file will contain twelve months (three interviews) of data from the 1984 SIPP panel, with the individual as the principal analytic unit. The sample of cases to be weighted for this file will be only those persons with three completed interviews. Those sample persons with only one or two interviews will be treated as nonrespondents. Their reported data will help to define nonresponse adjustment classes.

Since the first microdata longitudinal file contains only persons responding to all three interviews, the nonresponse adjustment issue is virtually the same as for the cross-section case. There are, however, two nonresponse adjustment factors applied to the initial sampling weights. See Kobilarcik and Singh (1986). The first adjustment factor accounts for households classified as noninterviews in the first interview wave. The second factor accounts for persons who did not supply all three interviews.

For the first adjustment factor, only those household variables available at the first interview can be used. Adjustment factors are calculated separately within cells defined by the following variables:

- a. Census Region
- b. Residence (metropolitan, non-metropolitan)
- c. Race of reference person
- d. Tenure (own, rent)
- e. Household size

The second set of adjustment factors is implemented on a person basis. The factors are calculated within cells defined by the following characteristics:

- a. Monthly household income
- b. Program participation status of the person's household
- c. Labor force status
- d. Race
- e. Years of school completed
- f. Type of assets of person's household

Cells are collapsed whenever they do not contain thirty sample persons or the nonresponse adjustment factor exceeds 2.

As the survey progresses, more sophisticated methods of adjusting for longitudinal nonresponse will be developed which make use of the data provided for partial respondents (i.e., for sample persons that provide some, but not all, of the interview waves requested). It is not obvious how to treat the partial response cases. Data gaps associated with persons who miss one or more interviews can be viewed as either person nonresponse, and typically handled by weighting adjustments, or as item nonresponse, usually handled by some type of imputation method. For example, one might consider an individual with a (R,NR,R) pattern as a case of item nonresponse since the missing interview is bounded on both sides by completed interviews; but one might consider an individual with an (NR,R,NR) pattern as total unit nonresponse, treating it the same as (NR,NR,NR). However, we need to recognize that even in the case of the response pattern (R,NR,R) for an individual, four kinds of response patterns are still possible at the item level. Thus, many options can be considered when developing nonresponse compensation procedures for the SIPP longitudinal data base. This issue is discussed by Kalton (1986) and by Kalton, Lepkowski, and Lin (1985).

3.2 SIPP Research Activities

There are two areas where work has recently begun which should aid future decisions concerning nonresponse adjustments. First, the SIPP questionnaire, beginning during the fourth interview, contains a "Missing Wave" section. This section uses a short series of questions on labor force participation, income sources, and asset ownership/nonownership for respondents in the current wave who did not respond in the preceding wave. Respondents who miss two or more consecutive interviews are not eligible to complete the "Missing Wave" section. By emphasizing data collection at the expense of minor reporting burden, the person nonresponse problem can be reduced to an item nonresponse problem. An evaluation of the quality of the retrospective data will be necessary prior to using these data.

The second area of work concerns general strategies in the treatment of person-wave nonresponse in the SIPP. Graham Kalton and his colleagues at the Survey Research Center will (1) compare longitudinal imputation and weighting strategies for handling person-wave nonresponse, (2) evaluate imputation and weighting models in terms of the analysis of change across waves and aggregation across waves, and (3) develop preliminary criteria for the choice of method for treating person-wave nonresponse. A discussion of these and other issues which will be studied can be found in Kalton (1986), and Kalton and Miller (1986).

Finally, there are several other research topics for which work is planned. These include: (1) quantifying the selection of variables used for determining weighting classes; (2) assessing the robustness of the survey estimates on the population and selected subgroups under different nonresponse compensation procedures, and different weighting class cell collapsing

strategies; (3) investigating the potential for making separate nonresponse adjustments by type of noninterview; (4) investigating the effect of deleting reported survey data to simplify the nature of the SIPP missing data problem; and (5) evaluating the longitudinal nonresponse compensation procedures adopted for the first SIPP longitudinal research file.

4. UNIT NONRESPONSE PROCEDURES FOR ECONOMIC CENSUSES AND SURVEYS

The Bureau of the Census carries out six economic censuses every five years, the most recent ones covering 1982. These six economic censuses are identified by the following trade areas:

- (1) Retail Trade
- (2) Wholesale Trade
- (3) Service Industries
- (4) Manufactures
- (5) Mineral Industries
- (6) Construction

In addition to the economic censuses, the Census Bureau carries out the Census of Governments and the Census of Agriculture. Though not part of the economic censuses, they are conducted during the same years as the economic censuses for processing efficiencies and to allow for data linkage. In nearly all of these economic areas the Census Bureau also carries out a number of monthly, quarterly, and annual surveys.

Like the demographic areas, there is some unit nonresponse for all of the economic censuses and surveys. In most cases, missing data are imputed based on (a) previous responses provided by the nonrespondent, (b) data from administrative records, and (c) relationships established between various data items. Rather than reporting the percent of units not responding, the level of nonresponse for an economic census or survey is usually given as the percent of one or more item totals that are imputed. These percents will be referred to as imputation rates.

Explanations of the unit nonresponse methods used for five of the six economic censuses are given in Section 4.1. Section 4.2 addresses unit nonresponse procedures for three economic surveys, and Section 4.3 covers such procedures for the Census of Agriculture. Research and evaluation activities with regard to nonresponse procedures for economic censuses and surveys are discussed in Section 4.4.

More detailed explanations of the nonresponse procedures used in these censuses and several related surveys are given by Bailey, Chapman and Kasprzyk (1985).

4.1 The Economic Censuses

The frame for the economic censuses is the Standard Statistical Establishment List (SSEL), a computer file maintained by the Census Bureau. The SSEL is comprised of all employer establishments reported by multi-unit employer companies in the Census Bureau's Company Organization Survey (COS) and all single-unit companies that filed a tax form with IRS. The COS is an annual survey of multi-unit companies. Companies that have at least 50 employees are surveyed each year, while companies with fewer than 50 employees are surveyed every three years. Each company in the COS is sent a list of the establishments it reported most recently in the survey and asked to update the list. They are also asked to provide, for each establishment, employee counts for the first quarter of the previous year and total payroll

for the previous year. For the economic censuses, each establishment on the SSEL, except small single-unit establishments, is sent a census questionnaire (via its company) designed for its standard industrial classification (SIC) code.

Although there are many similarities among the unit nonresponse procedures used in the six trade areas, some important differences exist. In the following description of the unit nonresponse adjustment procedures used for five of the economic censuses, the trade areas that use essentially the same procedure will be grouped together as follows:

- (a) Retail trade, wholesale trade, services
- (b) Manufactures, mineral industries

4.1.1. Retail Trade, Wholesale Trade, Service Industries

These three parts of the economic censuses are often referred to collectively as the business census. For these trade areas, data for the census year are collected on sales receipts, employment, and payroll. The imputation rate for sales/receipts varies from 10 to 15 percent for retail and wholesale trade and is about 20 percent for service industries.

For any establishment that does not provide the census data, responses are generally imputed using tax form information available from the Internal Revenue Service (IRS). For payroll information, the IRS has four quarters of data available for each employer identification (EI) number from tax forms. A company may have one or more EI numbers. Payroll data for a particular company are obtained by adding up the payroll figures for all EI numbers used by the company. First quarter employment counts are also available by EI number from IRS records and can be aggregated to the company level. For sales/receipts, various IRS forms are used depending on whether the nonresponding company is a sole proprietorship, partnership, or corporation.

The imputation procedure is complicated by the difference between the census enumeration unit and the IRS tax unit. For the business census, the unit of enumeration is the establishment (i.e., a single location). However, the tax unit for the IRS is an EI number. There may be one or more establishments reporting under the same EI number. If a nonresponding company has only one location (i.e., is a single-unit company), then it will have only one EI number and imputation is straightforward. However, for a multi-unit nonresponding company imputation is more complex since, in general, IRS data will not be available for each establishment. In such a case, the company structure is determined first by referring to the SSEL to obtain a list of all establishments contained in a company and all EI numbers used by the company. The total for a company for each data item is obtained by adding the item across all EI numbers used by the company, as discussed above. The company total is distributed to establishments by prorating the total based on the most recent data available for the company from an annual or monthly survey. If no data are available, an equal proration is used. If there is nonresponse for only a portion of the establishments in a multi-unit company, data for the nonresponding establishment are imputed based on prior year relationships.

4.1.2 Manufactures, Mineral Industries

In these two economic censuses, general information is obtained on the number of employees, hours worked, and on production levels by four-digit standard industrial classification (SIC) codes. Imputation rates vary from about 10 to 15 percent. The unit

nonresponse procedure used depends on the type of company that did not respond (i.e., single-unit or multi-unit) and on whether or not a previous year's record is available. Thus, there are four types of nonresponse cases that occur. The method of treating nonresponse for these four cases follows:

- (1) Single-unit company, previous year data are available from the Annual Survey of Manufactures.

In this case annual payroll is obtained from IRS tax forms and compared to the previous payroll total reported. The percent change from the previous period is determined. This percent change is applied to all data items in the previous record to obtain an imputed current record, except for employment and value of shipments whenever these are available from IRS.

- (2) Single-unit company, no previous year data are available.

In this case, sets of ratios are developed between census items within each four-digit SIC, with payroll as the "seed." That is, the relationships are developed in such a way that all items can be imputed from these relationships either directly or indirectly if a payroll figure is obtained. The specific relationships are derived from historic data reported by the respondents in the same industry. Then the (seed) value of payroll is obtained from IRS tax records and all other items are imputed from the relationships derived.

- (3) Establishment in a multi-unit company, previous year data are available for the establishment.

First, for each four-digit SIC, an aggregate growth factor between the previous and current period is developed from external sources for each of the following key items: payroll, employment, change in inventory, and change in capital expenditures. These four growth factors are applied to the appropriate prior year data items for each establishment to obtain imputed responses for the current period. These four imputed items are then used as "seeds" to impute other items.

- (4) Establishment in a multi-unit company, no previous year data are available for the establishment.

In this case, basic data on payroll and employment are obtained for each establishment from the SSEL discussed earlier in Section 4.1. As indicated, the SSEL obtains data on employment and payroll obtained for all establishments included in the COS. Then, using the SSEL data as a base, the data record for each establishment is imputed from relationships developed between the SSEL data items and the other census items. This procedure is analogous to that used in case (2) above.

4.2 Economic Surveys

The Census Bureau conducts a large number of monthly, quarterly, and annual economic surveys in addition to the economic censuses. In particular, most of the six census trade areas have monthly or annual surveys. The unit nonresponse procedures used for the Monthly Retail Trade Survey and the Truck Inventory and Use Survey are described below. The unit nonresponse adjustment procedure used for the Annual Survey of Manufactures (ASM) is not described here since it is virtually the same as that used for the Census of Manufactures, described in Section 4.1.2. Imputation rates for the ASM vary from 5 to 10 percent.

4.2.1 Monthly Retail Trade

The Monthly Retail Trade Survey includes about 30,000 reporting units: about 3,000 selected with certainty and 27,000 selected on a probability basis. The certainty cases are surveyed each month, while a third of the noncertainty cases are surveyed each month. This provides a monthly mailing of about 12,000 reporting units. For a multi-unit company in the survey, a subsample of the establishments in the company is selected for inclusion. Monthly retail sales is the only item enumerated in the survey. The imputation rate for retail sales is about 11 percent.

If a single-unit certainty company or a sample establishment in a multi-unit certainty company does not report for a given month, a value for sales is imputed from the previous month's figure by multiplying it by a "ratio of identicals." This adjustment ratio is derived by dividing the weighted sum of the current month sales by the weighted sum of the previous month sales for all establishments in the same adjustment cell for which sales were reported for both the current and previous months. Adjustment cells are generally defined by the first three digits (or four digits in a few cases) of the SIC code, by type of establishment (i.e., whether or not it belongs to a large multi-unit firm), and by sales size class. The weight used for each reporting unit used in computing the ratio of identicals is the inverse of the probability of selection of the reporting unit.

If a multi-unit certainty company does not report sales for any of its establishments, the sales values are imputed for each establishment and for the entire company as in the previous case: applying the ratio of identicals for the appropriate adjustment cell to the previous month sales figures. If such a company does report current monthly sales for the entire company, the imputed establishment responses are ratio adjusted to be consistent with the reported total for the entire company.

For noncertainty companies, imputation for missing sales data is carried out in a way similar to that used for certainty cases, except that an extra step is required since noncertainty companies report every three months. The first step is to impute the previous month's sales for a nonrespondent based on the response provided three months ago. This is done by multiplying the sales reported three months ago by a ratio of identicals based on the weighted sum of sales during the previous month and the weighted sum of sales three months ago (cell by cell). Once the previous month sales are imputed, the current month sales is generated from the imputed value for the previous month using the same method described for certainty cases.

If a nonrespondent is in the survey for the first time, the previous month's sales (if it's a certainty case) or the sales figure three months earlier (if it's a noncertainty case) is imputed from the sales reported in the most recent census, if available. If the nonrespondent was not in the most recent census, then it would be a birth case for which two months of sales data generally would have been provided at the time the company was added to the frame. This data would be seasonally adjusted and then inflated to an annual-based figure. The imputation would then be carried out as though a census sales figure had been available for the nonrespondent.

4.2.2 Truck Inventory and Use Survey (TIUS)

The TIUS is conducted every five years and provides data on the physical and operational characteristics of trucks nationwide. These characteristics include type of trailer (vehicle configuration), kinds of products carried, type of gasoline used, and annual miles driven. The universe for the survey consists of the truck registrations from all 50 states and the District of Columbia. The sample size is about 120,000 truck registrations. About 75 percent of the trucks selected for the survey respond.

Adjustments for unit nonresponse are made by “weighting up” the respondents to the total sample, separately within weighting classes. The weighting classes are taken to be the sample strata which consist of cross-classifications by state and body type (5 categories). The nonresponse weight adjustment is based on the number of trucks; within each class (stratum), the initial weight of each respondent is multiplied by the ratio of the number of trucks in the stratum to the sum of the initial weights of the respondents in the stratum.

Of the economic surveys investigated, the TIUS is the only one that uses a weight adjustment procedure to account for unit nonresponse. With other economic surveys, alternate sources of basic information are generally available to “build” a record for a nonrespondent.

4.3 Census of Agriculture

The census of agriculture provides data relating to the Nation’s farming, ranching, and related activities. It is the leading source of agricultural statistics and the only source of consistent, comparable data about agriculture at the county, State, and national levels.

The task of nonresponse adjustment for the census of agriculture is made complex by the fact that the SSEL cannot be as effectively used as it is in the other economic areas. The agricultural census mailing list is constructed by combining several overlapping sources. The resultant frame may contain some duplication and always contains some nonfarm entities. Thus, the nonresponse methodology must first identify, or estimate, the extent to which an adjustment is needed before it can take place.

For the 1982 census, nonrespondents were designated as large or small based on whether their expected sales were above or below \$100,000. A 100% telephone follow-up was conducted for all of the large nonrespondents. The small nonrespondents were then stratified based on other mail list characteristics. A sample of these units was followed up by mail and telephone to obtain estimates, by strata within states, of the percent of nonrespondents which were actually farms. These estimates were then used, along with data on in-scope percents of *respondents* by county, to make estimates of the number of nonrespondent farms at the county level for each stratum. The weights of a randomly selected sample of respondents by county, consistent with the estimated number of nonresponding farms, were then inflated by two. All other respondents retained their weight of one.

4.4 Research Activities for Nonresponse Adjustments in Economic Surveys

Probably the most important source of information for unit nonresponse imputation in economic surveys is IRS data from tax forms. Some differences between the IRS figures and those collected in the economic census may arise because of differences in definitions, forms, or the data collection procedures used. A study by Dyke (1984) compared administrative (IRS) data used to impute sales/receipts, payroll, and employment in the 1977 business census with corresponding responses obtained in a follow-up sample of nonrespondents. In general, he found that the survey values reported in the follow-up survey exceeded those obtained from administrative sources. The sizes of the differences varied by item. Also, the differences were more pronounced for multi-unit establishments. Additional comparisons of this type are needed. If systematic differences are identified, adjustment factors to apply to IRS figures may be developed.

For several of the censuses and surveys, a “ratio of identicals” is calculated and used to obtain a factor to apply to a previous-period figure to obtain an imputed value for the current period. It is possible that this ratio computed among *all* sample cases that reported in both periods may not apply very well to the nonrespondents for some items. Bailey (1986) looked at alternatives to using ratios of identicals for imputing missing values such as linear regression and quadratic regression, using various sets of independent variables.

With many of the economic unit nonresponse imputation methods, the sample cases – both respondents and nonrespondents – are placed into cells prior to computing (a) some type of ratio between current and prior periods for an item or (b) some type of relationship between the survey items and the basic items: payroll, employment, and receipts. A research project to investigate alternate choices of cell definition for the Monthly Retail Trade Survey was recently completed by Huang (1986). She found that for some SIC’s an alternate procedure of defining cells reduces the mean square error (MSE) of estimated sales substantially. In addition, she compared the current method of imputing – using ratios of identicals – to three alternate methods with respect to bias and MSE. The current method was evaluated as the second best procedure. However, she concluded that the slight gains of the optimum procedure may not be worth the additional requirements associated with using it.

5. IMPUTATION FOR EARNINGS IN THE CURRENT POPULATION SURVEY

5.1 The Hierarchical Hot Deck

The Current Population Survey (CPS) is a Census Bureau ongoing monthly survey of about 60,000 U.S. households per month. The CPS, sponsored by the Bureau of Labor Statistics, primarily collects labor force and employment information. Each March, the CPS administers an income supplement as part of the survey questionnaire. About 11-12% of the sample members do not respond to the income questions. Therefore, a special procedure, referred to as the “hierarchical hot deck,” has been developed to impute for missing responses.

With the hierarchical hot deck, missing earnings values are inserted from the response record of another sample unit – a donor. The goal in selecting a donor is to find one with survey characteristics similar to those of the item nonrespondent. The first step in the process of finding suitable donors is to partition the entire sample, excluding total noninterview cases, into cells based on multi-way classifications of a number of survey characteristics. Within each cell a list is made of the respondents and nonrespondents for a given item. Donors from the list of respondents are assigned to the nonrespondents systematically, with a random start. If there are more nonrespondents than there are respondents in a cell for a given item, the responses of some, or perhaps all, of the respondents in the cell will be used more than once. In some cells, there may be one or more nonrespondents but no respondents for an item.

To avoid the problem of having nonrespondents with no donors available, the process of defining cells and selecting donors for the item nonrespondents is carried out several times. At each stage, fewer cells are defined than were defined for the previous stage. For the final stage the number of cells defined is small enough so that it is certain that there will be donors available in each cell. The cells defined at successive stages are formed by collapsing the cells used at the previous stage. Each item nonrespondent will have one or more donors assigned. The donor used to obtain an imputed value will be the one identified at the earliest stage.

The major advantage of this hierarchical procedure is that a very large number of cells can be defined at the first stage, due to the backup stages used. Whenever a donor is found at the first stage, the item nonrespondent and donor will be matched on a large number of survey characteristics. In such cases there should be a good chance that an adequate imputation is made. In other cases the item nonrespondents and donors will be matched on fewer characteristics. This hierarchical procedure tries to pick donors in a way that maximizes the number of matched relevant survey characteristics.

For a more detailed description of this of this procedure, see Welniak and Coder (1980), Oh and Scheuren (1980a), or David, Little, Samuhel, and Triest (1986, Section 2).

5.2 Evaluation of the CPS Hierarchical Hot Deck

There have been some evaluation studies of the CPS Hot Deck: Welniak and Coder (1980); Oh and Scheuren (1980a and 1980b); Lillard, Smith, and Welch (1982); and David *et al.* (1986). One of the weaknesses noted of the CPS hot deck is that donor values may be used repeatedly, resulting in variance increases. The procedure could be modified to avoid using donor values more than once or twice; however, this change has not been made. The CPS hot deck procedure is based on the assumption that the distribution of responses for a survey variable is the same for respondents and nonrespondents in the same cell – the ignorability assumption.

David *et al.* (1986) developed several model-based alternatives to the CPS hot deck and evaluated them and the CPS hot deck with respect to mean absolute and mean relative error. These evaluations were based on a CPS-IRS matched file. In creating this file, an attempt was made to match the March 1981 CPS file to the IRS tax records for 1980. Despite the hot deck's apparent limitations, the CPS hot deck had a lower mean absolute and mean relative error than did the model-based alternatives. However, the models were developed for only 10% of the full CPS sample used to develop the hot deck procedure.

6. SUMMARY AND AREAS OF FUTURE STUDY

In this paper an attempt has been made, primarily through examples, to describe the current approaches being taken to nonresponse adjustments in the U.S. Census Bureau's censuses and surveys. Emphasis has been placed on the need for additional empirical and theoretical studies in both the demographic and economic areas in order to provide more objective guidelines (a) to design nonresponse compensation procedures and (b) to measure the effects of nonresponse on survey results for a variety of survey conditions.

Some of the research called for in this paper is already underway but more will be needed. For example, to what extent can available ancillary data be used in conjunction with modeling and data analysis procedures to identify the key functional relationships needed to provide a "reasonably" accurate description of the response/nonresponse structure applicable to a given survey?

In general, adjusting for nonresponse is just one of several steps taken to reduce the variance and bias of survey results. The degree to which these other steps aid in reducing the impact of nonresponse is an area for further research. Moreover, there should be continued efforts in support of research on recurring issues such as the impact of unit nonresponse weights and item nonresponse imputation on complex variance estimators, model approaches to determining appropriate adjustment factors, and the effectiveness of combining various types of nonresponse adjustment techniques.

ACKNOWLEDGEMENTS

The authors are very appreciative of the useful information received on the nonresponse adjustment procedures used for many of the U.S. Census Bureau's Censuses and Surveys.

This information was provided by personnel in several Census Bureau divisions, including Agriculture Division, Business Division, Construction Statistics Division, Economic Surveys Division, Governments Division, Industry Division, Statistical Methods Division, and the Statistical Research Division.

The authors are grateful to Dr. Fritz Scheuren for reviewing the draft and providing many helpful comments.

The authors are also indebted to Hazel Beaton, Alice Bell, and Valerie Howard for the diligence and patience they displayed in typing the manuscript.

REFERENCES

- ANDERSON, H. (1978). On nonresponse bias and response probabilities. *Scandinavian Journal of Statistics*, 6, 107-112.
- BAILEY, L. (1986). A study of alternative imputation techniques for surveys in the Current Industrial Reports. Internal Census Bureau Report, December 24.
- BAILEY, L., CHAPMAN, D.W., and KASPRZYK, D. (1985). Nonresponse adjustment procedures at the Census Bureau: A Review. *Proceedings of the Bureau of the Census First Annual Research Conference*, 421-444.
- DAVID, M., LITTLE, R.J.A., SAMUHEL, M.E., and TRIEST, R.K. (1986). Methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.
- DYKE, T.C. (1984). Evaluation of the use of administrative record data for establishments which were non-respondents to the 1977 Census of Wholesale Trade, Retail Trade, or Selected Services. Internal report: Statistical Research Division Report Series, No. Census/SRD/RR-84/08, U.S. Bureau of the Census.
- HANSON, R. (1978). The Current Population Survey: Design and Methodology. Technical Paper No. 40, Washington, D.C.: U.S. Bureau of the Census, pp. 55-59.
- HUANG, E.T. (1986). Comparison of different imputation procedures in the Monthly Retail Trade Survey. To appear in the *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.
- KALTON, G., and LEPKOWSKI, J., and LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP research panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 372-377.
- KALTON, G., McMILLEN, D. and KASPRZYK, D. (1986). Nonsampling error issues in the Survey of Income and Program Participation. *Proceedings of the Bureau of the Census Second Annual Research Conference*, 147-164.
- KALTON, G., and MILLER, M. (1986). Effects of Adjustments for Wave Nonresponse on Panel Survey Estimates. To appear in the *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- KOBILARCIK, E.L., and SINGH, R.P., (1986). SIPP: Longitudinal estimation for persons' characteristics. To appear in the *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- LILLARD, L., SMITH, J.P., and WELCH, F. (1982). What do we really know about wages: The importance of non-reporting and census imputation. *Journal of Political Economy*, 94, 489-506.
- LITTLE, R.J.A. (1986). Missing data in Census Bureau Surveys. *Proceedings of the Bureau of the Census Second Annual Research Conference*, 442-454.

- LITTLE, R.J.A., and SAMUHEL, M.E. (1983). Imputation models on the propensity to respond. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 415-420.
- NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985). An overview of the Survey of Income and Program Participation: Update 1. SIPP Working Paper Series No. 8401, U.S. Bureau of the Census.
- OH, H.L., and SCHEUREN, F.J. (1980a). Estimating the variance impact of missing CPS income data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 408-415.
- OH, H.L., and SCHEUREN, F.J. (1980b). Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data. *Proceeding of the Survey Research Methods Section, American Statistical Association*, 416-420.
- OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*, Vol. 2, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 143-184.
- PALMER, S., and JONES, C. (1967). A look at alternate imputation procedures for CPS noninterview. *Proceedings of the Social Statistics Section, American Statistical Association*, 73-80.
- PAUL, E.C., and LAWES, M. (1982). Characteristics of respondent and nonrespondent households in the Canadian Labour Force Survey. *Survey Methodology*, 8, 48-85.
- POLITZ, A., and SIMMONS, W. (1949). An attempt to get the 'Not-At-Homes' into the sample without callbacks. *Journal of the American Statistical Association*, 44, 9-31.
- ROSENBAUM, P., and RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- THOMSEN, I., and SIRLING, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. In *Incomplete Data in Sample Surveys*, Vol. 3, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 25-59.
- U.S. Department of Commerce, Bureau of the Census (1977). National Crime Survey, national sample, survey documentation. U.S. Bureau of the Census Report.
- U.S. Department of Commerce, Bureau of the Census (1983). Cross-sectional weighting specifications for the first wave of the 1984 panel of the Survey of Income and Program Participation (SIPP). Internal U.S. Bureau of the Census Memorandum from C. Jones to T. Walsh, November 25.
- U.S. Department of Commerce, Bureau of the Census (1984a). Economic characteristics of households in the United States: Third Quarter 1983. *Current Population Reports, Series P-70, No. 1*, Washington, D.C.: U.S. Government Printing Office.
- U.S. Department of Commerce, Bureau of the Census (1984b). 1984 SIPP first wave weighting-first stage estimate factors and specifications for collapsing noninterview adjustment calls. Internal U.S. Bureau of the Census Memorandum from C. Jones to T. Walsh, February 16.
- U.S. Department of Commerce, Bureau of the Census (1984c). SIPP weighting: subsequent wave cross-sectional - revised. Internal U.S. Bureau of the Census Memorandum from C. Jones to T. Walsh, October 12.
- WELNIAK, E.J., and CODER, J.F. (1980). A measure of the bias in the March CPS earnings imputation scheme. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 421-425.
- YCAS, M., and LININGER, C. (1981). The income survey development program: Design features and initial findings. *Social Security Bulletin*, Vol. 44, No. 11, November.

Hot Deck Imputation Procedure Applied to a Double Sampling Design

SUSAN HINKINS and FRITZ SCHEUREN¹

ABSTRACT

From an annual sample of U.S. corporate tax returns, the U.S. Internal Revenue Service provides estimates of population and subpopulation totals for several hundred financial items. The basic sample design is highly stratified and fairly complex. Starting with the 1981 and 1982 samples, the design was altered to include a double sampling procedure. This was motivated by the need for better allocation of resources, in an environment of shrinking budgets. Items not observed in the subsample are predicted, using a modified hot deck imputation procedure. The present paper describes the design, estimation, and evaluation of the effects of the new procedure.

KEY WORDS: Double sampling; Hot deck; Imputation.

1. INTRODUCTION

When the U.S. Internal Revenue Service (IRS) is mentioned, the first words to cross one's mind may not be "sample surveys." But every April, those of you from the U.S. take part in at least one of our administrative "surveys" and file an individual income tax return. We sample this administrative data annually for statistical purposes. Another of our major programs is an annual sample of U.S. corporate tax returns; that is the sample survey discussed here.

The primary interest at a Symposium like this is in non-response or other undesirable missing data. Despite our extensive enforcement efforts, we at IRS also have such non-response problems. However, the present paper is concerned with a different type of missing data problem: missingness that is not unexpected, but is designed (see also, Strudler, Oh, and Scheuren 1986, for another example). We take the liberty of discussing these problems because we use techniques usually associated with non-response, e.g., hot deck imputation (Ford 1983). Our case allows an evaluation of the imputation procedure, since the underlying non-response mechanism is known.

Double sampling has been introduced in our corporate tax return sample in an effort to reduce costs with only a "tolerable" loss of information. Reweighting to account for the subsampling stage is a standard estimation approach in double sampling (e.g., Cochran 1977); however, in our application, we would have had to reweight almost on an item-by-item basis. This was judged unacceptable by our users, who require rectangular data sets. (For an analogous approach in a Canadian context, see Colledge *et al.* 1978.)

The imputation technique used - hot deck imputation - is procedurally simple. The need to discuss the application of such a relatively simple procedure may surprise theoreticians; but, as we will show, the problems of implementation within the setting of a large statistical operation are many.

¹ Susan Hinkins, Statistics of Income Division, Internal Revenue Service, P.O. Box 369, Bozeman, Montana 59771.
Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Avenue, N.W., Washington, DC 20224.

In the remainder of the present paper, we describe in some detail the double sampling procedure and the imputation technique employed. Preliminary results on the impact of these procedures are also presented and the last section contains our conclusions and future plans. A brief theoretical discussion of the estimators we are using and their properties is given in an Appendix.

2. DESCRIPTION OF THE SAMPLING PROCEDURES

An annual sample of U.S. corporate tax returns is used by IRS to estimate National totals of both tax and economic variables. For example, approximately three million corporate tax returns will be filed for 1985, and the IRS sample will contain over 90,000 of these returns. (In Canada, there are two separate corporate tax return samples, each designed to meet narrower purposes. The Revenue Canada Taxation sample (e.g., Burpee and McGrath 1982) was developed for tax policy simulation purposes. The Statistics Canada sample (e.g., Ambrose 1985) is intended primarily to estimate economic aggregates. It is our belief that separate designs in the U.S., but not entirely separate processing systems, could lead to improvements in efficiency over the current procedures; however, the work done (Clickner *et al.* 1984) indicates that the problem is quite difficult and progress has been slow.)

The annual estimates obtained are for the entire corporate population and for subpopulations, usually defined by industrial activity and size. The underlying population is highly skewed. For most variables, a small proportion of the population accounts for a substantial fraction of the total dollar amount. Examples for 1982 corporations are given in Exhibit 1.

A highly stratified sample design is used; small corporations are selected with small probability and large corporations are selected with certainty (Jones and McMahon 1984). The strata are defined by industrial classification and the size of the corporation (i.e., in terms of assets and net income). Selection probabilities for each stratum are determined by employing a modified form of Neyman allocation. Almost all of the returns in the 100% strata (returns selected with certainty) have total assets of \$50 million or more. A form of post-stratified raking ratio estimation is used to weight the sample results (Leszcz, Oh, and Scheuren 1983).

Retrieving the information from each sampled return is a time-consuming and expensive process. Over 600 items may be retrieved from a return, and these items are not simply extracted; they are also carefully checked and redistributed to compensate for taxpayer reporting variations. The complete process is referred to as "editing the return". The cost of "editing" varies by degree of complexity. It may take only twenty-five minutes to edit a fairly simple return but as long as a week to edit a really complicated one. The quality of the editing is vital to our estimates, as these checks reduce, but do not eliminate reporting inconsistencies.

Exhibit 1
Degree of Concentration of Selected Corporate Variables

Selected Items	Assets Under \$50 Million	Assets \$50 Million or more
Number of Returns	99.6%	0.4%
Total Assets	16.3	83.7
Total Receipts	39.3	60.7
Total Income Tax	25.9	74.1

Source: Internal Revenue Service, 1985.

Indeed, nonsampling error is a serious concern in the data "editing" process, particularly for the largest corporations. In order to spend proportionately more resources on reducing the nonsampling error for the large returns, we introduced stratified double sampling for the smaller returns; specifically, certain data items were retrieved on only a subsample of the returns (i.e., a subset of returns with assets under \$50 million). Although this change would increase the error for some variables on the small returns, we expected that the procedure would have little adverse effect on the estimates of national totals, or on the subdomain estimates of primary interest to our major users. There were two main reasons for this conjecture:

- As already noted, corporate *returns* with total assets of \$50 million or more were not subject to the extra sampling step.
- The information loss due to the subsampling was reduced by the choice of the *items* or variables to be subject to subsampling.

By and large, as will be shown, the results obtained so far confirm our expectations.

Items Selected for Subsampling

When certain miscellaneous items on a return are nonzero, the taxpayer must attach a schedule providing additional information. For example, if the item "Other Income" is nonzero, the corporation must describe what was included under this category. The schedules are attached on separate sheets of paper and have no standard form or length. The process of editing a schedule has several parts: finding the schedule, deciding whether the taxpayer included appropriate amounts in "Other Income", and making changes if there are errors.

Beginning with the tax year 1981 corporate program, the statistical editing of data from the tax return was done in stages, and certain items were initially transcribed for statistical use directly from the return. Employing automatic tests, items or schedules could then be "flagged" for abstraction or further scrutiny in later stages (Cys *et al.* 1982). This new strategy allowed us to:

- Retain original taxpayer information as reported so that the amount of editing change could be evaluated. Prior to the 1981 sample, we had no information regarding the extent of the adjustments being made by editing. The editors only recorded the final result. (See Powell and Stubbs 1981.)
- Decide whether or not to review a particular schedule based on the initial information transcribed. (Again, prior to the 1981 program, editors were, of course, required to completely edit all schedules.)

For the 1981 and 1982 corporate programs, seven items and their associated schedules were picked for subsampling: schedules for Other Income, Other Deductions, Other Costs of Goods Sold, Other Current Assets, Other (Noncurrent) Assets, Other Current Liabilities and Other (Noncurrent) Liabilities.

The reported amounts on a corporate return may be modified substantially as a result of the editing. For example, consider the "Other Income" schedule shown in Exhibit 2. The original amounts (in column 1) are observed initially for every return. The variables being subsampled are changes that would be made if the Other Income schedule were edited (column 2). In this hypothetical case, we have an original Other Income amount of \$1,600, which, when examined by the editor, could be reclassified as including \$900 from Business Receipts, \$300 in Rents and \$400 that really belongs in Other Income. The variables of interest are, of course, the final ("corrected") amounts for each item.

Before implementing the new processing system, an experiment was run comparing the amount of time it took to do the reduced, initial transcription and the amount of time it took to do the complete editing (reading all schedules). As expected, the reduced edit was

Exhibit 2
Illustration of Editing Other Income

Income Type	Original Amounts(\$)	Change Amount(\$)	Final Amounts(\$)
Other Income	1,600	- 1,200	400
Receipts	500	+ 900	1,400
Rents	0	+ 300	300
Interest	700	0	700

significantly faster (and therefore, cheaper). Considerable resources could be saved by sub-sampling. (Conservatively, we extrapolated 1981 cost savings of at least \$300,000, assuming only limited use of the subsampling technique.)

Double Sampling

We are now ready to describe the basic two-dimensional stratification chosen for our double sampling. The returns are stratified into “crucial” returns (Group A) versus the remaining returns (Group B). “Crucial” returns include all returns with total assets of \$50 million or more, thereby including the important “large” returns and most returns selected into the sample with certainty. In addition, crucial returns should include corporations of any size for which the likelihood of an editing change was high. What we want, obviously, is a sub-sampling plan that has us edit all schedules that have a high probability of a change (especially a large change) and lets us subsample the rest.

In an attempt to predict which schedules are likely to change, a record is included in Group A if the original amount in Other Income, to continue our illustration, is unusually large compared to the amount in Total Income.

Also, since we do not want to impute large amounts, cases where Other Income is above a certain dollar value should be included in Group A, as well. (Unfortunately, this was done only indirectly.) By inference, Group B is supposed to include only small returns which we believe are likely to have little or no change made as a result of editing. (See Barker *et al.* 1982, for details.)

For the crucial returns in Group A, all variables (items) are always completely observed. Only returns in Group B are subject to the subsampling of the seven schedules mentioned earlier. Even for Group B returns, the original amounts for all items are always recorded; therefore, some information is obtained for every item. The information not obtained for some records in Group B is the change due to editing a schedule. It is these changes that are being imputed using the procedure described in the next section. Not all variables are affected by the subsampling. For example, of the 600 items picked up for the 1981 corporation program, only 56 were in any way affected by the double sampling; however, of the approximately 100 major income and balance sheet items, nearly one half could be affected.

3. THE IMPUTATION PROCEDURE

The missing information (i.e., changes from editing) in Group B was imputed using a hot deck procedure within adjustment cells. A record with schedules to be imputed was matched to a donor record, in the same adjustment cell, with these same schedules edited. (The formation of adjustment cells is described later in this section.)

In 1981, the subsampling rate was 10% for the returns subjected to subsampling: one out of ten was selected systematically for editing (these were the hot deck "donors") and the other nine were left to be imputed. In 1982, the subsampling rate was kept at 10% for non-financial returns (trade, manufacturing, etc.) but was raised to 20% for financial returns (banks, insurance companies, etc.)

Within an adjustment cell, the number of returns, n' , can be divided into the number of donors, n'' , and the number of imputes, $n' - n''$. Because of the small subsampling rate, the number of donors is almost always smaller than the number of imputes. In particular, let $n' - n'' = rn'' + t$ where r and t are nonnegative integers and $0 \leq t < n''$. Then the hot deck procedure selects all n'' donors r times, and selects the remaining t units by simple random sampling without replacement.

To continue our illustration, recall that the item of interest is Z , the final "corrected" amount for Other Income; Z can be written as $Z = X - Y$, where X is the original taxpayer amount in Other Income and Y is the change made due to editing the Other Income schedule. It is only the change, Y , that is unobserved and must be estimated for a subset of the returns in Group B.

If we simply employ a conventional hot deck procedure and estimate the unobserved y_i value, on record i , with the observed value y_j from donor record j , then the resulting estimate of the final value z_i may not satisfy the edit checks. For example, assume the donor record had \$30,000 originally as Other Income, and \$15,000 was removed when the schedule was edited. Suppose that on the record to be imputed, the original amount in Other Income is \$10,000, then the imputed change of \$15,000 would result in a negative estimate for other income:

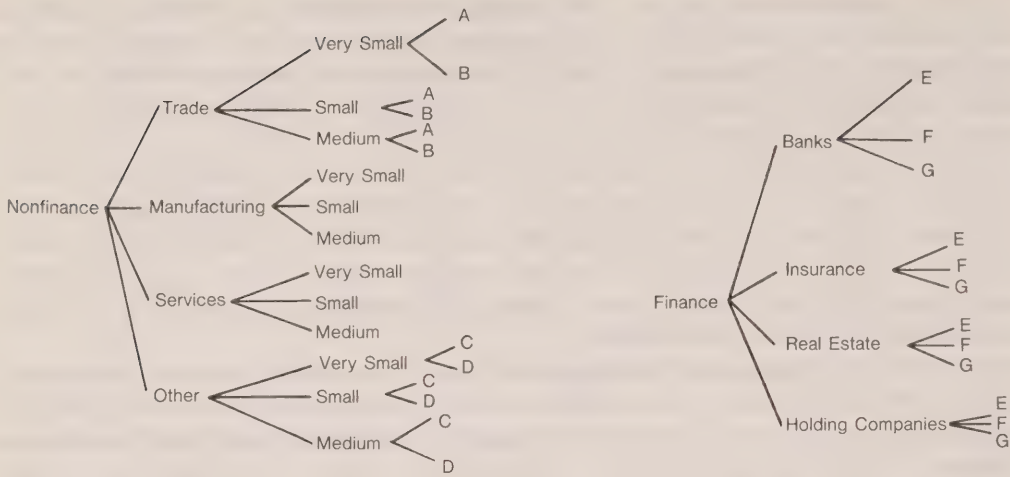
$$\hat{z}_i = x_i - \hat{y}_i = 10,000 - 15,000 = -5,000.$$

Since the amount for Other Income must be nonnegative, edit checks would fail and additional adjustments would have to be made to the record. (See Sande 1982, for a general discussion of this problem.) Since the original amount is always observed, it seemed more reasonable to "hot deck" the relative change $R = Y/X$ rather than the actual change Y . In this example, since the donor record had one half of the amount in Other Income removed after reading the schedule, then 1/2 should be removed on the imputed record. The estimated final amount in Other Income is then

$$\hat{z}_i = x_i - \hat{y}_i = 10,000 - (1/2)10,000 = +5,000.$$

In addition to satisfying the edit checks, we expected the ratio procedure to reduce the variance of our estimates relative to the basic hot deck approach; however, the variance of the estimator is not analytically tractable and must be measured empirically. We have not yet verified in our corporation application the smaller variance that we conjecture; but simulation results do support the approach we have taken. However, by introducing the ratio, our estimators are now biased. We conjectured that the biases would be small and in fact they were, for the most part, as we shall show.

The model associated with our imputation procedure is based on the definition of the double sampling strata being used and on the definition of the adjustment cells. Several constructive steps were taken to make the approach reasonable. In the initial stratification, an attempt was made to subsample only those records that were likely to have no changes or only small changes. Also, the adjustment cells were *subjectively* chosen to be homogeneous with respect to the magnitude of the relative editing change that might be made. In particular,



The coded tree branches above correspond to the following:
A = Retail, B = Wholesale, C = Transportation and Utilities, D = Other, E = Very Small,
F = Small, G = Medium.

Figure 1. Hierarchy of Ratio Hot Deck Adjustment Cells

the adjustment cells are defined in terms of industrial classification, corporation size and the pattern of items present on the return. There were thirty categories defined by various industrial and size criteria (see Figure 1). In addition, sixteen item patterns were treated separately, defined by the presence/absence of Other Income (2 classes), the presence/absence of either Other Deductions or Other Costs of Goods Sold (2 classes), Other Current Assets or Other Assets (2 classes) and, finally, Other Current Liabilities or Other Liabilities (2 classes). The maximum number of adjustment cells was $30 \times 16 = 480$.

For each item pattern, a hierarchical structure was developed so that collapsing could be done when there were an insufficient number of donors for use in the imputation (see Figure 1). The first division is into financial returns (banks, insurance companies, etc.) versus non-financial records; cells are not collapsed across this division. The next levels of the hierarchy separate cases according to fairly broad industrial classes and according to the size of the corporation, in terms of assets and net income. Recall that the largest corporations are not subject to subsampling and, so, should not need imputation; hence, broad industrial and size groups seemed sufficient.

The quality of our estimation depends on how much collapsing takes place. In 1981, we had 36,586 returns with at least one schedule to impute, and 3,989 donors. For the non-financial returns we never collapsed across the major industrial classification, and, in fact, we always had some size distinction. Many cells were not combined at all, but maintained the maximum detail possible. In contrast, for financial returns the size variable was often lost by combining all cells, and major industries were sometimes combined (Hinkins 1983). For one pattern, all financial returns were combined into the same cell.

Based on our 1981 experience, several changes were made in the 1982 double sampling design:

- Due to the extensive collapsing of cells for financial returns in 1981, the subsampling rate for small financial returns was doubled to improve the estimates (from 10% to 20%, as noted earlier).

Table 1
Selected Statistics on Hot Deck Ratio Imputation, 1981-1982

Item	Tax Year 1981		Tax Year 1982	
	Financial	Non-financial	Financial	Non-financial
NUMBER				
Donors	908	3,081	1,806	4,697
Imputes	7,912	28,674	10,719	43,477
Adjustment Cells	113	238	142	260
DONOR CELL SIZE				
Average	8	13	13	18
Maximum	68	58	126	98
Minimum	1	1	2	2
DONOR-TO-IMPUTE RATIOS				
Average	.11	.11	.17	.11
Maximum	1.00	.25	2.00	.28
Minimum	.05	.05	.05	.05

Note: For 1982, cell sizes of 2 donors each were required in order to make possible the calculation of the variance.

- In 1981, the double sampling procedure was not applied across the entire sample, but was restricted to certain processing centers. Other processing centers collected all information, as before. In 1982, the procedure was applied across the whole sample. The relative number of records in 1982 with some items imputed was 63 percent, compared to 40 percent in 1981.
- In order to estimate the hot deck imputation variance (Oh and Scheuren 1980; Rubin and Schenker 1986), an additional restriction was imposed on the 1982 design, in that we required that there be at least two donors in each adjustment cell. (See Table 1.)

In 1982, there were 54,196 records to be imputed from 6,503 donors, and there was considerably less collapsing of adjustment cells (Hinkins 1984). In particular, for financial records, 94 percent of the records imputed in 1982 were in adjustment cells defined with some size distinction, compared to 75 percent in 1981. Table 1 provides a selection of other statistics on the operation of the 1981 and 1982 systems.

4. INITIAL EVALUATION OF BIAS

The evaluation of the 1982 double sampling system is still underway, but some initial results are available on the potential biasing effects of the imputation. Bias should be small if R , the ratio of the editing change to the original amount, is always small, or if R is constant within adjustment cells. We have taken the approach of looking for the "worst" cases of bias by looking for examples where R is neither small nor constant. We confine attention to only two variables: Other Income and Business Receipts.

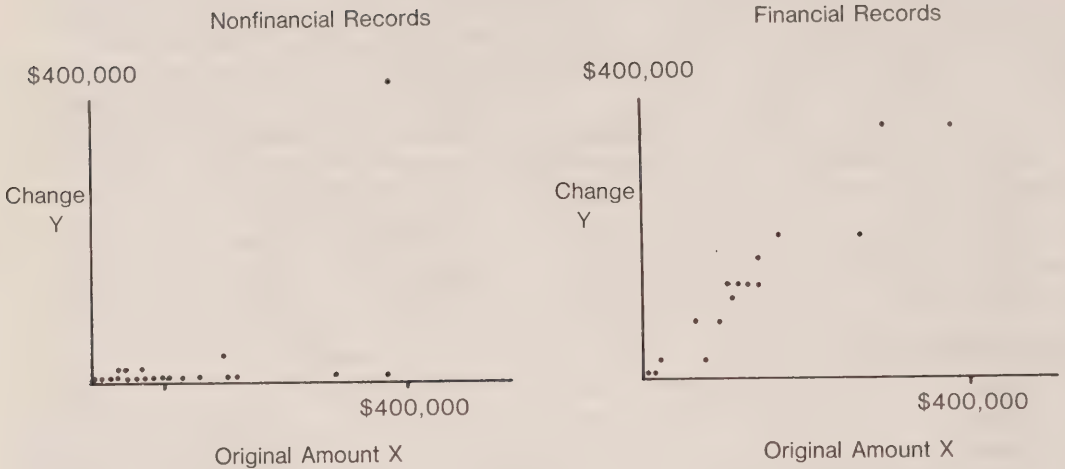


Figure 2. Changes in Other Income: Group B Donors only

Unbiased Model

The ratio bias in the hot deck imputation we are using would be zero if the relationship $Y = RX$ were to hold for all members of each adjustment cell chosen. An overall plot of the data might be useful, to look at the degree to which this model holds for Other Income. In Figure 2, therefore, we have plotted the Group B donors separately for financial and non-financial corporations. There is a distinct difference between these two categories. Nonfinancial returns are much less likely to change; in 1982, 14 percent of the nonfinancial donors had a change made to Other Income, compared to 59 percent of the financial records. Also, for financial returns at least, it looks as if the model $E(Y) = RX$ might be appropriate. Further work along these lines is intended, but the scatterplot encourages us to believe that, by and large, existing biases would be small.

Actual Bias Measures

Table 2 provides relative bias measures for selected worst case industries. These are shown for all returns in that industry and returns with assets under \$25 million (i.e., for corporations likely to be most affected by the new procedures). Of the items changed in the double sampling the Other Income schedule showed some of the largest values of R and the most disperse distributions of R . The greatest change as a result of editing Other Income was made in the Business Receipt amount. It should be noted that the bias estimates in Table 2 are subject to considerable sampling error (Czajka 1986). Except for the very smallest amounts, however, it is conjectured that the estimates shown probably have the correct sign and are of the appropriate order of magnitude.

These examples indicate that within small subpopulations, there can be noticeable bias effects. However, even within a major industry, selected for its potential problems, the bias across all sizes is relatively small.

Table 2
Estimated Relative Biases for Business Receipts and Other Income
by Selected Minor Industries, 1982

Selected Minor Industries	Business Receipts		Other Income	
	All Returns	Assets Under \$25 Million	All Returns	Assets Under \$25 Million
(Biases as percent of applicable total)				
WHOLESALE TRADE				
Machinery, Equipment and Supplies	-1.40	-2.6	0.4	0.6
Miscellaneous Trade	-0.30	-0.5	-1.3	-2.4
RETAIL TRADE				
Auto Dealers and Service Stations	-0.30	-0.5	3.3	4.6
FINANCE AND INSURANCE				
Banking	-0.02	-0.7	0.1	2.4
Credit Agencies Except Banks	-0.50	-2.2	-0.9	-9.0
Insurance Agents	-0.60	-0.7	1.2	2.3

Note: All calculations are based on design-weighted estimates of the biases involved. The industries were selected to represent worst case examples.

Czajka's results (1986) indicate that for global estimates (across all industries), the bias effect of the imputation is small (less than 1% in all cases; considerably less than .05% in most cases).

There is no question that some of the biases in Table 2 appear large and warrant concern; however, it is important to realize that the overall effect on the root mean square error of the bias is small for all returns, generally 5% or less. These results give us strong evidence that the procedures employed did little or no harm to the data needed by our users; that, however, is not to say that major improvements, like those envisioned for 1985 and 1986, should not be made.

5. FUTURE PLANS AND SUMMARY

Double sampling and imputation were not used for the 1983 and 1984 samples because of processing constraints. They will be used again starting with the 1985 sample. As part of reinstituting the imputing process, we are planning to make several changes:

- It will no longer be necessary to initially transcribe certain items for statistical purposes before subjecting the records to double sampling. The fields needed are now being obtained directly from the IRS revenue processing system, so they are available before we begin reading and editing the tax return; thus, before editors first look at a return, we can designate whether or not they should review certain schedules. This makes the use of stratified double sampling even more appealing; the savings should increase.
- However, because of the new processing system, only three schedules are now available for subsampling. The schedules for 1985 are Other Income, Other Deductions and Other Costs of Goods Sold; the remaining four schedules used in 1981 and 1982 had to be dropped from the subsampling design.

- Despite the modest success of the 1981 and 1982 procedures, changes will be made for 1985 in the imputation methods. For example, the current definition of the adjustment cells could be improved, and separate imputation depending on the pattern of items represented needs to be reconsidered. The possible use of predictive mean matching within adjustment cells also bears examination (Little 1986). For 1986, refinements in the subsampling plan will need to be looked at too.
- Finally, we would like to base our estimates, in some way, on previous years' data, so as to be able to impute missing information earlier in the processing. In order to minimize the collapsing of adjustment cells, the 1981 and 1982 imputation processing had to wait for all records to be available. This delayed production by several weeks. We could avoid this problem by further increasing the number of donors; but, the editing of more records has the obvious disadvantage of increasing costs. On the other hand, by basing our approach in part on the previous year's data, we might not only improve the estimation, but also allow the imputation calculations to be done in the mainstream of processing.

Overall Summary

In this paper, we have described the reasons we had for making major changes in our statistical processing of corporate returns:

- The traditional complete data estimate was rejected in favor of double sampling because of cost considerations.
- The usual double sampling estimator (reweighting the complete data) was rejected because it did not result in a rectangular data set.
- A conventional hot deck approach was rejected because the resulting estimates could fail the edit checks.

Instead, the relative change was estimated using ratio hot deck imputation within adjustment cells.

We conjectured that because the double sampling procedure was restricted to a subset of the "small" corporations, the estimates of interest to our major users should be virtually unaffected; indeed, these estimates could even be improved, by better allocating our resources to validate and correct the records of the larger corporations. Our results so far largely vindicate these conjectures.

Compared to the traditional complete data estimator, the use of double sampling and hot deck imputation increased the mean square error of estimates in two ways; bias was introduced, and the variance of the estimator was increased. Our preliminary results indicate that there could be a significant bias effect for some estimates; however, the examples were chosen because they appeared to be cases where the hot deck ratio method would be weakest. Even so, the estimated overall effect of the procedure on the root mean square error appears relatively small. Looking at the increase in variance, the largest component is usually due to the decrease in sample size (double sampling). This increase in variance also turned out to be relatively small, since only one component of the final amount (the change) is imputed; the variance of the original values appears to dominate the variance of the changes.

In conclusion, while there are improvements to make, we feel encouraged to continue with our current double sample design and imputation technique. Perhaps at another Conference of this type we will be able to report on the further results of our research.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge the considerable help they have received from the staff members in the SOI Division, whose day-to-day responsibilities are covered by the material presented here. We would also like to thank David W. Chapman and John L. Czajka for their many constructive comments, especially in clarifying our exposition. We, of course, accept full responsibility for any remaining obscurities or errors.

APPENDIX: SOME BASIC THEORY

This appendix provides some technical details on the double sampling procedure as applied in our particular situation. We contrast several potential estimators for the double sampling design we chose. An overall summary of the bias and variance expressions for these different approaches is found in Table A.

For this discussion, we ignore the underlying stratified sample design and act as if a simple random sample had been taken, or equivalently we consider estimates within a sampling stratum. To do otherwise would make the notation exceedingly complex, but would not change the main points we wish to make.

Let us again consider just one of the items subject to subsampling, namely Other Income as before. The variable of interest is Z , the final, corrected value of Other Income, and Z can be decomposed as

$$Z = X - Y,$$

where X = the original taxpayer (or revenue processing) value of Other Income,

Y = the change made to Other Income after reviewing the schedule.

The population values and parameters are indicated by upper-case letters and the sample statistics by lower case. The population parameters of interest are the finite population mean and variance, i.e.,

$$\bar{Z} = \sum Z_i / N = \bar{X} - \bar{Y},$$

$$S^2(Z) = \sum (Z_i - \bar{Z})^2 / (N - 1).$$

Complete Sample - Prior to the introduction of double sampling, the estimates were calculated from a complete sample of size n' , and the unbiased estimator of \bar{Z} was

$$\begin{aligned} \bar{z} &= \sum z_i / n' \\ &= \bar{x} - \bar{y}. \end{aligned}$$

Ignoring the finite population correction (N is large), the variance is

$$\text{Var}(\bar{z}) = S^2(Z) / n'.$$

Table A
Selected Properties of Alternative Estimators

Estimator	Bias	Variance	Satisfy Edit?
Complete Sample	0	$\text{Var}(\bar{z})$	Yes
Double Sample	0	$\text{Var}(\bar{z}) + c_1 S_B^2(Y)$	Yes
Hot Deck Amount (Y)	0 ^a	$\text{Var}(\bar{z}) + c_1(1 + c_2) S_B^2(Y)$	No
Ratio (R)	b_1	$\text{Var}(\bar{z}) + V_1$	Yes
Combined Ratio	b_2	$\text{Var}(\bar{z}) + V_2$	Yes

^a In general, the basic hot deck procedure is unbiased only when it results in final values that satisfy the edit checks.

In Table A, we use the properties of \bar{z} as a benchmark, to compare among alternative estimators.

Double Sampling Estimation – Using Cochran’s notation (Cochran 1977, 12.2), the original sample of size n' has now been stratified into the two groups A and B, with n_A' and n_B' units respectively. A subsample of size n_B is selected from group B. The original taxpayer amount X is recorded for all $n' = n_A' + n_B'$ records. The changes due to editing Other Income, Y , will be recorded for all n_A' units in group A and for the random subsample of n_B units in group B.

Since the double sampling procedure only applies to variable Y , within group B, the double sampling estimator of \bar{Z} is

$$\begin{aligned}\bar{z}_d &= \bar{x} - \bar{y}_d \\ &= \bar{x} - \left(\sum y_{Ai} + (n'_B/n_B) \sum y_{Bj} \right) / n'\end{aligned}$$

and \bar{z}_d is unbiased.

- Let N_B = number of population units falling in stratum B ,
- P_B = N_B/N , proportion of population falling in stratum B ,
- \bar{Y}_B = population mean in stratum B ,
- $S_B^2(Y)$ = $\Sigma (Y_{Bi} - \bar{Y}_B)^2 / (N_B - 1)$, $i = 1, 2, \dots, N_B$,
- $1/K$ = the subsampling proportion = n_B/n'_B .

If the sampling proportion, $1/K$, is assumed fixed (in our application, $1/K = .10$ or $.20$), it follows (Cochran 1977) that the unconditional variance of \bar{z}_d is, ignoring the fpc,

$$\begin{aligned}\text{Var}(\bar{z}_d) &= \text{Var}(\bar{z}) + c_1 S_B^2(Y), \\ &= [S^2(Z) + P_B(K - 1) S_B^2(Y)] / n',\end{aligned}$$

where $c_1 = P_B(K - 1) / n'$.

Therefore the price paid for the reduction in cost due to not editing every schedule, is the increase in variance due to double sampling. This increase in variance looks potentially damaging because K is large. However, recall that $Z = X - Y$, and the increase in variance is a function only of the variance of Y within subpopulation B. We expect $S^2(X)$ to dominate $S^2(Y)$, which should further dominate $S_B^2(Y)$, i.e.

$$S^2(X) \gg S^2(Y) \gg S_B^2(Y).$$

This is because the size of the variance is related to the mean value, and Y should be small compared to X . (For most items, we expect the amount misclassified to be small, compared to the original amount). Therefore we expect $S_B^2(Y)$ to be so much smaller than $S^2(Z)$ that $P_B(K - 1)S_B^2(Y)$ will still be relatively small compared to $S^2(Z)$, and so the increase in variance due to subsampling will be relatively small. This is not guaranteed, but Czajka's results bear this out, for most items (Czajka 1986).

Hot Deck Imputation – Hot deck imputation was used, within adjustment cells, to reconstruct a rectangular data set. In particular, a return with schedules to be imputed was matched to a donor in group B, in the same adjustment cell, with these same schedules edited.

Imputing the missing values of y with a hot deck procedure, using simple random sampling, further increases the variance over using the double sampling estimate (\bar{z}_d). However the additional increase in variance due to using hot deck imputation is small compared to the increase due to double sampling. This relative increase in variance due to imputing, denoted as c_2 in Table A, is bounded and in our case is small. (When $K \geq 2$, $c_2 \leq 0.125$. See, for example, Hansen, Hurwitz, and Madow 1953).

As discussed in the paper, there is a problem with using an ordinary hot deck approach. If we simply estimate the unobserved y_i value, on record i , with the observed value y_j from donor record j , then the resulting estimate of the final value z_i may not satisfy the edit checks. Additional corrections would have to be made to the record. Since the original amount is always observed, it seemed more reasonable to "hot deck" the relative change $R = Y/X$ rather than the actual change Y . In addition to satisfying the edit checks, we expected the ratio procedure to reduce the variance of our estimates relative to the basic hot deck approach; however the variance of our estimator is not analytically tractable and must be measured empirically. Also, by introducing the ratio, our estimators are now biased. We conjectured that the biases would be small and in fact they were, for the most part, as seen in Table 2. In practice, the hot deck imputation was done within adjustment cells, created by post-stratifying the records into what we hope are homogeneous cells. The effect of this post-stratification should be to reduce variance and bias effects, but that is dependent on our skill in defining the imputation cells (an area with ample room for additional work).

Ratio or Regression Estimation – We are also considering ratio (or regression) estimates within cells, instead of the hot deck estimates. For example, $\hat{z} = x_i - \hat{r} x_i$, where $\hat{r} = \bar{y}/\bar{x}$ is calculated within appropriate cells. Referring to Table A, the increase in variance, V_2 , using the ratio estimator could be approximated using the formulas for the ratio estimator (e.g., Cochran 1977). However, these formulas are large sample approximations, and our sample sizes are almost always quite small. (In this case, the sample size is the number of donors, n_B , in an adjustment cell.) Therefore, empirical results are needed here.

Similarly, the bias, b_2 , can be found using the results for ratio estimators. Unlike the hot deck ratio, the bias of the ratio estimator goes to zero as the sample size increases and in this sense the ratio estimator is more robust. In fact, the hot deck ratio estimator is unbiased only if the model $Y = \beta X$ is correct. (Of course, the bias of both estimators goes to zero as the fraction of missing data goes to zero). However, even if the model $Y = \beta X$ is incorrect, the ratio estimator is consistent.

There are of course many other options; multivariate regression models could be investigated. We are still in the early stages of this project and we certainly have our work cut out for us now and in the upcoming years.

REFERENCES

- AMBROSE, P. (1985). Tax year 1985 business finance (T2) sample selection: detailed statement of requirement. Statistics Canada (Unpublished).
- BARKER, D., HINKINS, S., and REHULA, V. (1982). 1981 corporation validation tests. Statistics of Income Division, Internal Revenue Service (Unpublished).
- BURPEE, J., and McGRATH, A. (1982). Micro-model of corporation taxation sample design and estimates. Statistical Services Division, Revenue Canada Taxation (Unpublished).
- CLICKNER, R.P., GALFOND, G.J., and THIBODEAU, L.A. (1984). Evaluation of the IRS corporate SOI sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 443-448.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: John Wiley and Sons, Inc.
- COLLEDGE, M., JOHNSON, J., PARE, R., and SANDE, I.G. (1978). Large scale imputation of survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 431-436. (See also the paper by S. Michaud in this issue.)
- CYS, K., HINKINS, S., and REHULA, V. (1982). Automatic and manual edits for corporation income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 443-448.
- CZAJKA, J. (1986). Imputation of selected items in corporate tax data: improving upon the earlier hot deck. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (in publication).
- FORD, B.L. (1983). An overview of hot deck procedures. In *Incomplete Data in Sample Surveys*, Volume 2 - Theory and Bibliographies (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 185-207.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, Vol. II. New York: John Wiley and Sons, Inc.
- HINKINS, S. (1983). Matrix sampling and the related imputation of corporate income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 427-433.
- HINKINS, S. (1984). Matrix sampling and the effects of using hot deck imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 415-420.
- JONES, H., and McMAHON, P. (1984). Sampling corporation income tax returns for statistics of income, 1951 to present. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 437-442.
- LESZCZ, M.R., OH, H.L., and SCHEUREN, F.J. (1983). Modified raking estimation in the Corporate SOI Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 434-438.

- LITTLE, R.J.A. (1986). Missing data in Census Bureau surveys. Presented at the Second Annual Census Research Conference, March 1986. To appear in the *Journal of Business and Economic Statistics*.
- OH, H.L., and SCHEUREN, F.J. (1980). Estimating the variance impact on missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 416-420.
- POWELL, W.T., and STUBBS, J.R. (1981). Using business master file data for statistics of income purposes. *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. 1., Washington, DC: Internal Revenue Service, 157-167. See, especially, the Appendix by Alan Freiden.
- RUBIN, D., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- SANDE, I.G., (1982). Imputation in surveys: coping with reality. *The American Statistician*, 36, 145-152.
- STRUDLER, M., OH, H.L., and SCHEUREN, F.J. (1986). Protection of taxpayer confidentiality with respect to the IRS Individual Tax Model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (in publication).

Comparison of Weighting and Imputation Methods for Estimating Unsampled Data

SYLVIE MICHAUD¹

ABSTRACT

The Canadian Census of Construction (COC) uses a complex plan for sampling small businesses (those having a gross income of less than \$750,000). Stratified samples are drawn from overlapping frames. Two subsamples are selected independently from one of the samples, and more detailed information is collected on the businesses in the subsamples. There are two possible methods of estimating totals for the variables collected in the subsamples. The first approach is to determine weights based on sampling rates. A number of different weights must be used. The second approach is to impute values to the businesses included in the sample but not in the subsamples. This approach creates a complete "rectangular" sample file, and a single weight may then be used to produce estimates for the population. This "large-scale imputation" technique is presently applied for the Census of Construction. The purpose of the study is to compare the figures obtained using various estimation techniques with the estimates produced by means of large-scale imputation.

KEY WORDS: Weighting; Large-scale imputation; Unsampled.

1. INTRODUCTION

The Census of Construction (COC) is an annual survey which attempts to estimate expenses in the construction field. Although it is called a "census", in fact only businesses having a gross income exceeding \$750,000 are surveyed. Various financial and non-financial data are collected by means of a long questionnaire mailed to these firms. For businesses with a gross income between \$10,000 and \$750,000, expenses are estimated from a sample of administrative data. First, two samples are selected independently from overlapping sample frames. Two subsamples are then drawn from one of the samples in order to obtain additional information.

Variables collected in the subsamples may be estimated in two different ways. The method currently used for the Census of Construction is to impute values for the businesses included in a sample, but not in a subsample. This creates a complete "rectangular" file, from which estimates for the overall population may be produced using only one weight. An alternative would be to calculate weights based on the probabilities of selection; these would have to be calculated separately for different subsets of data. The purpose of this study is to compare the estimates obtained by weighting with the estimates obtained by imputation.

The study was carried out on a population of unincorporated businesses only because, for fiscal year 1983, the sample selection strategies for unincorporated and incorporated businesses were different. The strategy used for corporations will be modified for fiscal 1984 to be equivalent to the strategy for unincorporated businesses. The strategy for unincorporated businesses was therefore examined. One hopes that the conclusions of this study will remain the same for incorporated businesses.

¹ S. Michaud, Business Survey Methods Division, Statistics Canada, 11th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

2. DESCRIPTION OF THE SAMPLING PLAN

As mentioned above, two independent samples are drawn from overlapping sample frames. The first is the prespecified sample selected for the Census of Construction; it is stratified by gross business income (GBI), province and 3-digit 1970 Standard Industrial Classification (SIC) code. The sample frame used is not completely up-to-date. It contains some “deaths”, i.e. businesses which are no longer within the scope of the COC for various reasons (a firm which no longer exists, is no longer engaged in a construction activity, or whose gross income is below \$10,000). Furthermore, the sample frame does not contain “births” or businesses which have changed activities and are now part of the construction industry. The second sample is a “cross-sectional” sample, selected independently by Revenue Canada from a complete database containing businesses in all SIC groups (not only construction). It is used to estimate “births”. This sample is stratified by Gross Business Income ranges. Figure 1 below illustrates the situation.

Two independent subsamples are selected from the units of the prespecified sample: a financial subsample and a subsample of “other characteristics” (OC). The OC subsample is drawn directly from the prespecified sample, while the financial subsample is selected using data transcribed from the sample (and so “deaths” are not subsampled). Further details concerning the sampling plan may be found in Giles (1983).

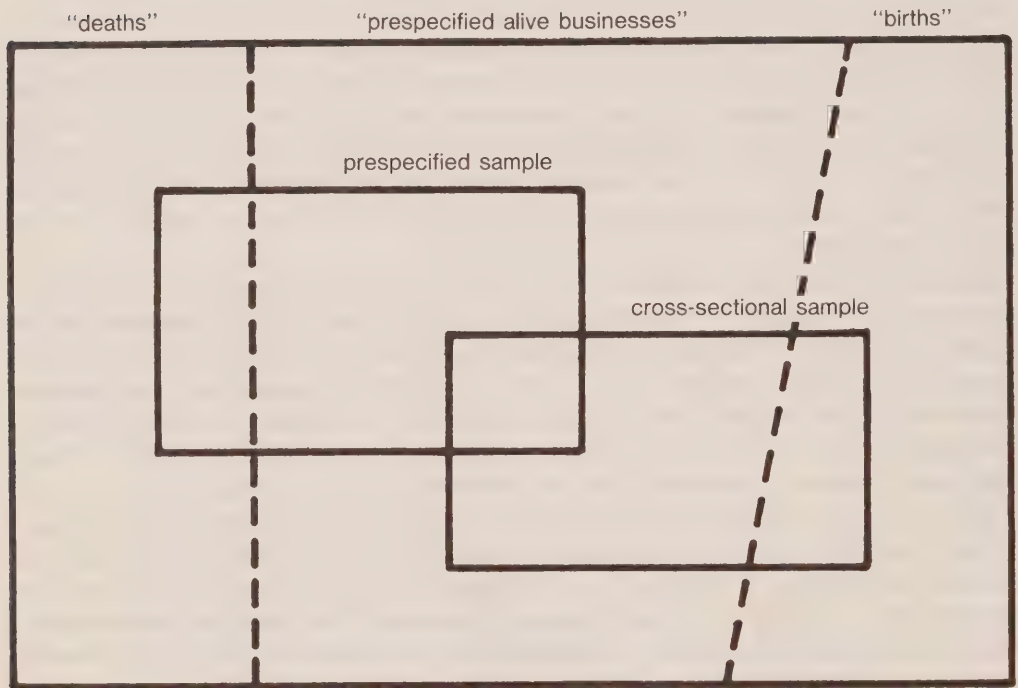


Figure 1. Representation of RC Sampling Plan

3. IMPUTATION TECHNIQUE

The COC uses a large-scale imputation technique to estimate the variables selected in a given subsample (i.e. values are imputed for each variable, for all records not selected in the subsamples). The imputation is carried out independently for each subsample. (The imputation is done in phases, and the imputation phases of the various subsamples are mutually independent and apply different techniques.) In each phase, the nearest neighbour is chosen from a subset of potential donor records, and is used to impute the variables which were not sampled.

The imputation is carried out differently for each subsample.

In the case of the financial subsample, the imputed value is the donor's value, adjusted by the ratio of an auxiliary variable which is available for both the donor and the candidate (the candidate being the record which is missing data to be imputed). (Note: The actual procedure is more complicated: the variables are imputed hierarchically and linear constraints are placed on the imputed values (the second variable is dependent on the value imputed to the first variable, etc.). Additional information on this procedure may be found in Philips and Emery (1976). A more detailed overview is also provided in Colledge *et al.* (1978)).

Suppose we use the following notation:

- Y : the variable of interest (known for the donors, to be imputed for the candidate)
- X : an auxiliary variable available for both the donor and the candidate
- c : denotes the candidate
- d : denotes the donor
- I : denotes an imputed value.

For the financial subsample variables, the imputed value Y_c^I is defined to be:

$$Y_c^I = Y_d \frac{X_c}{X_d}$$

For the OC subsample variables, the imputed value is simply the value on the donor record:

$$Y_c^I = Y_d$$

The imputation procedure produces a complete rectangular file (the records of all the businesses that were selected in one of the samples contain values for all the variables of the samples/subsamples). Sampling weights may then be used to generate estimates for the overall population.

The weight assigned to a given record is the inverse of the probability of it being selected into at least one of the samples. If we use the following notation:

- $P(\text{presp}_h)$: the probability of a record being selected in stratum h of the prespecified sample
- $P(\text{cross}_k)$: the probability of a record being selected in stratum k of the cross-sectional sample
- hk : cross-classification of records
- h : denotes the stratum of the prespecified sample
- k : denotes the stratum of the cross-sectional sample,

then the weight associated with each unit may be expressed as:

$$W_{hk}^{-1} = 1 - [1 - P(\text{presp}_h)] [1 - P(\text{cross}_k)]$$

Births and deaths cannot be cross-classified. Deaths have a zero weight $W_h = 0$ and the weight of a birth, W_k , is the inverse of the probability of being selected in stratum k of the cross-sectional sample. More details may be found in Bankier (1982).

Therefore, when the imputation technique is used, the estimator of the total is

$$\hat{Y} = \sum_{h,k} W_{hk} \sum_{j=1}^{n_{hk}} y_{jhk}^*$$

where $y_{jhk}^* = y_{jhk}$ if $j \in$ subsample

$= y_{jhk}^I$ if $j \notin$ subsample.

4. WEIGHTING TECHNIQUE

If a weighting technique were used to estimate subsample variables, there would be a number of possible estimators. The estimators are in the same form for both subsamples, but different weights are used.

The first estimator (\hat{Y}_1) would be based on the sampling plan used, adjusted for undercoverage of the population. In each of the SIC, PROV and GBI strata (Standard Industrial Classification, province, gross business income), a prespecified sample is selected. Once they have been transcribed (units sampled and still alive), the units are classified to two strata: "outside survey field" and "within survey field". The subsamples are chosen from the "within survey field" stratum. (We may assume that all the units in the "outside survey field" stratum have been subsampled and have a mean equal to zero.) The estimator contains a correction factor that compensates for undercoverage of the sample frame (calculated using information from the cross-sectional sample).

The second possible estimator (\hat{Y}_2) is a simplified version of the first estimator, \hat{Y}_1 . Instead of assuming a double sampling to determine "within survey field" and "outside survey field" units, we could assume that a prespecified stratified sample is selected from "within survey field" units. A subsample is selected from the prespecified sample. The estimator must once again be adjusted to take undercoverage into account. If the differences between the first and second estimator turn out to be insignificant, the second would be a better choice because it is simpler.

The third possible estimator (\hat{Y}_3) is an estimator based on data from the cross-sectional sample only. We could assume that the units selected in both the subsample and the cross-sectional sample are selected from the cross-sectional sample. The reasoning behind such an estimator is that the cross-sectional sample is drawn from a complete sample frame. However, since the subsamples are selected from the prespecified sample, and not from the cross-sectional sample, the size of the subsamples in the cross-sectional sample will be small.

Finally, a fourth estimator (\hat{Y}_4) could be obtained by supposing that the subsample is selected from the complete sample (prespecified sample + cross-sectional sample), and that the complete sample comes from multiple frames. This fourth estimator is the one that most closely resembles the estimator obtained after large-scale imputation. Indeed, both of these estimators assume that births and new businesses "react" like the rest of the population. The imputation procedure does not make any special adjustment for such businesses, and the weighted estimator is not stratified in such a way as to distinguish these units. In addition, both estimators take into account the fact that the sample comes from a number of frames. The same sampling weight is therefore used in both cases to produce data up to the population level.

As mentioned above, the variables collected in the financial subsample are adjusted by the ratio of an auxiliary variable during the imputation.

We could therefore propose another type of estimator for the variables collected in the financial subsample: a ratio estimator. The auxiliary variable used would be the same one used for the imputation. As is the case for the simple weighting, different estimators could be calculated.

The various estimators and their variances are described in mathematical terms in the Appendix.

5. RESULTS

In the study, four of the seven variables in the financial subsample were considered.

As for the subsample of other characteristics, eight variables are collected for all businesses, while other variables are available for certain SIC groups only. The study was therefore limited to these eight variables.

The variables in the financial subsample presented in this report are “ADD” (additions to fixed assets) and “RM” (repair and maintenance). For the OC subsample, results are given for the variable “PCON” (percentage of construction in a specific field). However, the PCON variable is not published directly, but is multiplied by total expenses to obtain expenses in a specific field: PEXP. This second variable was the one studied.

As mentioned earlier, the variables in the OC subsample are not adjusted by a ratio during the imputation procedure. The ratio estimators will therefore not apply to these variables.

Tables 1, 2 and 3 provide values for the different estimators and estimates of their respective variances, based on 1983 tax data for unincorporated businesses.

In the first place, we see that there are no significant differences between the first two estimators. (According to the predetermined definitions, the second estimator is a simplified version of the first one.)The simplified version will therefore be retained.

Table 1
Estimated Values of PEXP (%EXP*EXPCONS) and Standard Deviation of PEXP

	\hat{Y}_1	\hat{Y}_2	\hat{Y}_3	\hat{Y}_4	\hat{Y}_1
Estimate ($\times 10^{11}$)	3.44	3.43	3.96	3.66	3.70
Standard deviation ($\times 10^9$)	3.5	3.5	8.4	3.2	

Table 2
Estimated Values of ADD and Standard Deviation of ADD

	\hat{Y}_1	\hat{Y}_2	\hat{Y}_3	\hat{Y}_4	\hat{Y}_{Q2}	\hat{Y}_{Q3}	\hat{Y}_{Q4}	\hat{Y}_1
Estimate ($\times 10^8$)	2.08	2.10	2.14	1.84	7.82	5.06	5.2	1.4
Standard deviation ($\times 10^7$)	1.9	1.9	2.0	1.0	0.8	2.2	0.8	

Table 3
Estimated Values of RM and Standard Deviation of RM

	\hat{Y}_1	\hat{Y}_2	\hat{Y}_3	\hat{Y}_4	\hat{Y}_{Q2}	\hat{Y}_{Q3}	\hat{Y}_{Q4}	\hat{Y}_1
Estimate ($\times 10^8$)	1.5	1.5	1.43	1.55	0.9	1.63	1.67	1.75
Standard deviation ($\times 10^6$)	6.9	6.9	8.9	5.3	3.1	11.0	4.3	

In general, for the variables in the financial subsample, the imputation technique appears to yield results similar to those produced by the weighting method (\hat{Y}_4). The estimator obtained by considering only units drawn from the cross-sectional sample (\hat{Y}_3) seems more variable than the other estimators. This variability could be explained by the smaller number of units used to calculate this estimator. It should be pointed out that these comparisons are based only on an observed sample, and so the conclusions are somewhat limited. However, owing to the nature of the data (often percentages and subdivisions of activity in the construction field), which is relatively stable in the strata (3-digit 1970 SIC, province and GBI), it was considered unnecessary to analyse these variables in greater depth.

For the variables in the financial subsample, it was found that the estimators adjusted by the ratio do not always seem applicable (for example, the ADD variable). The estimates which they produce are extremely biased. One possible explanation is that the ADD variable and the auxiliary variable used have a high frequency of zero values. A “bad” sample in certain strata can thus inflate the estimates inordinately.

Some problems were also encountered with the imputation system (data imputed when they should not have been, data not imputed), which in certain instances may have affected the estimates obtained by the imputation method. Since the results were based on an observed sample only, and because it was difficult to estimate the impact of the system-related problems, it was decided that a simulation would be done.

6. SIMULATION

The simulation was carried out using a data subset, namely those businesses that had been selected in the financial subsample (all of the variables studied are present for this data subset). Then an attempt was made to apply a simplified version of the technique used by the Census of Construction. A stratified sample was selected, using sampling rates similar to those of the survey. The variables of the financial subsample, for the data not selected in the sample, were considered as missing, and then imputed by the system. The sample selection process and the imputation were repeated thirty times.

Estimates were produced, allowing us to compare the results obtained by summing the non-imputed and imputed data with the estimates produced using sampling weights equal to the inverse of the sampling rate. Since the value for the population is known, the bias and the variance of the estimates were calculated. The results for the ADD and RM variables are shown in Tables 4 and 5.

For the ADD variable, the value produced by ratio estimation differs significantly from the estimates obtained by imputation or by weighting. The bias of the estimate is also significantly not null. For the RM variable, all the estimators are equivalent (equal variances, bias not significant at a 5% level, estimates not significantly different).

Table 4
ADD Estimates Obtained by Simulation

	Population	Weighting	Ratio	Imputation
Estimate ($\times 10^7$)	1.41	1.43	1.24	1.41
Standard deviation ($\times 10^5$)		1.11	.85	1.15
Bias ($\times 10^5$)		.22	-1.73	-0.07

Table 5
RM Estimates Obtained by Simulation

	Population	Weighting	Ratio	Imputation
Estimate ($\times 10^7$)	1.06	1.06	1.07	1.04
Standard deviation ($\times 10^5$)		4.52	4.11	4.87
Bias ($\times 10^5$)		-0.07	-0.95	-1.38

7. CONCLUSIONS

According to the study results, there do not appear to be significant differences between the large-scale imputation technique and the weighting technique, for the variables in the other characteristics subsample. This was foreseeable, inasmuch as the variables studied seem to be relatively stable within each stratum.

The conclusions for the variables in the financial subsample are based on the results of the simulation. These seem to indicate that the estimates obtained by weighting by the inverse of the probability of selection are comparable to the estimates obtained from large-scale imputation.

The ratio estimator does not appear appropriate for the ADD variable (or for the other variables analysed, but not discussed in this report). Continuation of the study will try to determine whether a regression estimator would be more appropriate, and to evaluate the impact of the imputation on the variable correlation structure.

ACKNOWLEDGEMENT

The author would like to thank P. Giles for his comments and suggestions given during this study, and also the referees for their helpful comments.

APPENDIX

The following notation may be used for the proposed estimators:

- h : stratum of the prespecified sample
- k : stratum of the cross-sectional sample
- N_h : size of the "prespecified" population in stratum h
- \hat{N}_{1h} : size of the "prespecified" population with "alive businesses (within the scope of the survey) in stratum h (estimated)

- \hat{N}_{2h} : size of the “prespecified” population with businesses “outside the scope of the survey” in stratum h (estimated)
 \hat{N}_k : size of the population in stratum k , estimated using information from the cross-sectional sample
 \hat{N}'_k : size of the population in stratum k , estimated using information from both samples (multiple frames)
 n_h : number of units sampled in stratum h of the prespecified sample
 \hat{n}_{1h} : number of units sampled and transcribed in stratum h of the prespecified sample
 \hat{n}'_k : number of units sampled and transcribed in stratum k
 \hat{m}_{1h} : number of units subsampled from among “alive” businesses in stratum h
 y : variable of one of the subsamples
 x : auxiliary variable available for all units of the samples
 s^2_{yh} : estimate of the variance of y for the units of the subsample in stratum h
 s^2_{xh} : estimate of the variance of x for the units of the subsample in stratum h
 s_{yxh} : estimate of the covariance of x and y in stratum h .

$$\begin{aligned}
 \text{i)} \quad \hat{Y}_1 &= \left(\frac{\hat{N}_{1 \text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1 \text{ pre-spec.}}} \right) \sum_h \frac{N_h}{n_h} \frac{\hat{n}_{1h}}{\hat{m}_{1h}} \sum_{j=1}^{\hat{m}_{1h}} Y_{hj} \\
 V(\hat{Y}_1) &\approx \left(\frac{\hat{N}_{1 \text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1 \text{ pre-spec.}}} \right)^2 \sum_h N_h n_h \left(\frac{N_h - 1}{n_h - 1} \right) \\
 &\times \left[W_{1h} S_h^2 \left(\frac{1}{\gamma_h} - \frac{1}{N_h} \right) + \frac{G_h}{n_h} S_h^2 \left(\frac{W_{1h}}{N_h} - \frac{1}{\gamma_h} \right) + \frac{G_h}{n_h} W_{1h} (1 - W_{1h})^2 \bar{y}_h^2 \right]
 \end{aligned}$$

where $G_h = \left(\frac{N_h - n_h}{N_h - 1} \right)$, $\gamma_h = n_h \frac{\hat{m}_{1h}}{\hat{n}_{1h}}$, and $W_{1h} = \frac{\hat{n}_{1h}}{n_h}$.

$$\begin{aligned}
 \text{ii)} \quad \hat{Y}_2 &= \left(\frac{\hat{N}_{1 \text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1 \text{ pre-spec.}}} \right) \sum_h \frac{\hat{N}_{1h}}{\hat{m}_{1h}} \sum_{j=1}^{\hat{m}_{1h}} y_{hj} \\
 V(\hat{Y}_2) &\approx \left(\frac{\hat{N}_{1 \text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1 \text{ pre-spec.}}} \right)^2 \sum_h \frac{\hat{N}_{1h}}{\hat{m}_{1h}} (\hat{N}_{1h} - \hat{m}_{1h}) s_{yh}^2
 \end{aligned}$$

$$\begin{aligned}
 \text{iii)} \quad \hat{Y}_3 &= \sum_k \frac{\hat{N}_k}{\hat{m}_{1k}} \sum_j Y_{kj} \\
 V(\hat{Y}_3) &= \sum_k \hat{N}_k \left(\frac{\hat{N}_k - \hat{m}_{1k}}{\hat{m}_{1k}} \right) S_{yk}^2
 \end{aligned}$$

$$\text{iv) } \hat{Y}_4 = \sum_k \frac{\hat{N}'_k}{\hat{m}_{1k}} \sum_{j=1}^{\hat{m}_{1k}} y_{kj}$$

$$V(\hat{Y}_4) = \sum_k \hat{N}'_k \left(\frac{\hat{N}'_k - \hat{m}_{1k}}{\hat{m}_{1k}} \right) s_{yk}^2.$$

Ratio estimators may be calculated and, like simple estimators, they may take on different forms, depending on the hypotheses postulated. For example, the ratio estimator corresponding to estimator 4 would be:

$$\hat{Y}_{Q4} = \sum_k \hat{N}'_k \bar{Y}_{\text{sub}k} \frac{\bar{X}_{\text{samp}k}}{\bar{X}_{\text{sub}k}}$$

where $\bar{X}_{\text{samp}k}$ is the mean of variable X for the units selected in the complete sample, which are in stratum k

$\bar{X}_{\text{sub}k}$ is the mean of X for the units selected in the subsample, which are in stratum k

$\bar{Y}_{\text{sub}k}$ is the mean of variable Y in stratum k of the subsample.

$$V(\hat{Y}_{Q4}) = \sum_k (\hat{N}'_k)^2 \left(\frac{1}{\hat{m}_{1k}} - \frac{1}{\hat{n}'_{1k}} \right) \left[s_{yk}^2 + \hat{R}_k^2 s_{xk}^2 - 2\hat{R}_k s_{yxk} + \left(\frac{1}{\hat{n}'_{1k}} - \frac{1}{\hat{N}'_k} \right) s_{yk}^2 \right]$$

$$\text{where } \hat{R}_k = \frac{\bar{Y}_{\text{sub}k}}{\bar{X}_{\text{sub}k}}.$$

REFERENCES

- BANKIER, M., (1982). Variance formula for an estimator based on any number of independant stratified samples of which some are Poisson samples. Technical document, Business Survey Methods Division, Statistics Canada.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.
- COLLEDGE, M.L., JOHNSTON, J.H., PARÉ, R., and SANDE, I.G. (1978). Large scale imputation of survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 721-726.
- GILES, P. (1983). Construction division: Census of Construction. Technical document, Business Survey Methods Division, Statistics Canada.
- PHILIPS, J.L., and EMERY, D. (1976), FIBCO documentation. Technical document, Systems Development Division, Statistics Canada.
- RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

A Regression Approach to Estimation in the Presence of Nonresponse

CARL ERIK SÄRNDAL¹

ABSTRACT

In the presence of unit nonresponse, two types of variables can sometimes be observed for units in the "intended" sample s , namely, (a) variables used to estimate the response mechanism (the response probabilities), (b) variables (here called co-variables) that explain the variable of interest, in the usual regression theory sense. This paper, based on Särndal and Swensson (1985 a, b), discusses nonresponse adjusted estimators with and without explicit involvement of co-variables. We conclude that the presence of strong co-variables in an estimator induces several favourable properties. Among other things, estimators making use of co-variables are considerably more resistant to nonresponse bias. We discuss the calculation of standard error and valid confidence intervals for estimators involving co-variables. The structure of the standard error is examined and discussed.

KEY WORDS: Response mechanism; Adjustment group method; Co-variate; Robustness.

1. INTRODUCTION

We consider a finite population $U = \{1, \dots, k, \dots, N\}$ from which a sample s of size n is drawn with a sampling design under which the k -th unit has the (strictly positive) probability π_k of being selected. The sampling weight associated with the k -th unit is thus π_k^{-1} . We may admit a complex sampling design, not necessarily self-weighting, for example, a three-stage design with stratified selection of primary units. The probability under the design of jointly including the units k and l is denoted π_{kl} ($\pi_{kl} > 0$ for all $k \neq l$, and π_{kk} is interpreted as equal to π_k).

Given s , a certain unit nonresponse is assumed to occur. The responding subset of s is denoted by r , its size by m . The variable of interest, y , is observed for $k \in r$ only. To counteract the biasing effects of the nonresponse, we assume for the purpose of this paper that the widely used adjustment group method is employed: the sample s is subdivided into H groups $s_1, \dots, s_h, \dots, s_H$ of respective sizes $n_1, \dots, n_h, \dots, n_H$. The response set r is correspondingly divided into the subsets $r_1, \dots, r_h, \dots, r_H$, of respective sizes $m_1, \dots, m_h, \dots, m_H$. The response rate in group h is denoted $f_h = m_h/n_h$. The method calls for attaching (in addition to the sampling weight) the "adjustment weight" f_h^{-1} to an observation coming from group h . (The sizes and the composition of the adjustment groups at the population level are here assumed unknown.) We have:

$$n = \sum_{h=1}^H n_h; \quad m = \sum_{h=1}^H m_h.$$

¹ Carl Erik Särndal, Department of Mathematics and Statistics, University of Montreal, Montreal, Quebec, Canada, H3C 3J7.

Let $t = \sum_U y_k$ be the unknown population total to be estimated. (If A is an arbitrary set of units, we shall systematically write $\sum_A y_k$ for $\sum_{k \in A} y_k$.) The usual adjustment class estimator of t then becomes

$$\hat{t} = \sum_{h=1}^H f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}. \tag{1.1}$$

The adjustment group method is motivated theoretically by an assumption that units within the same group respond with the same (unknown) response probability. (More formally, this is expressed as Model *A* in Section 3 below.) The method clearly requires that group identity can be determined for each unit $k \in s$. The (categorical) variables that permit this grouping can thus be regarded as variables used for the estimation of an underlying response mechanism.

A different category of variable may be observable for each $k \in s$, namely, variables that explain y , in the ordinary regression theory sense. These variables will be termed co-variables. When incorporated in the estimator, such variables will not only reduce variance but also make the estimator more resistant to nonresponse bias. (They are not auxiliary variables in the usual sense of this term, since they are available not for the entire population U but only for the intended sample s .)

We shall thus keep a firm distinction in this paper between two types of variables observed for $k \in s$, those that are used to estimate the response mechanism, and those that explain the target variable y . Little (1983), in presenting a general framework for data with nonresponse, distinguishes several types of variables. One attempt to describe our situation in terms of Little's setup would be to say that the set of complete item variables in Little's terminology are, in our case, further subdivided into one subset of variables used to model the nonresponse mechanism, and another subset (the co-variables) serving as explanatory variables for the incomplete item variable y . Our approach to inference is that of "quasi-randomization" (Oh and Scheuren 1983), where "quasi" refers to the fact that the non-response selection phase must be modelled, whereas the sample selection phase is controlled by the sampler.

**2. SOME SIMPLE NONRESPONSE ADJUSTED ESTIMATORS
OF THE POPULATION TOTAL**

A slight development of the often seen formula (1.1) leads to a (generally somewhat "better") alternative in which the sampling weights π_k^{-1} can be said to be more fully used:

$$\hat{t}_{\text{EXP}} = \left(\sum_s \frac{1}{\pi_k} \right) \frac{\sum_{h=1}^H f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}}{\sum_{h=1}^H f_h^{-1} \sum_{r_h} \frac{1}{\pi_k}}.$$

The formula (which becomes identical to (1.1) for a self-weighting design) can be written as an expansion of the response set mean:

$$\hat{t}_{\text{EXP}} = \hat{N} \bar{y}_r,$$

namely, if we let the expansion factor be $\hat{N} = \sum_s 1/\pi_k$, and

$$\tilde{y}_r = \frac{\sum_{h=1}^H f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}}{\sum_{h=1}^H f_h^{-1} \sum_{r_h} \frac{1}{\pi_k}}. \quad (2.1)$$

The symbol tilde will be used to indicate a properly weighted mean statistic. The "tilde mean" \tilde{y}_r , being a response set mean, is calculated by attaching to the k -th unit the multiplicative weight:

$$\text{sample weight} \times \text{non response adjustment weight} = \pi_k^{-1} f_h^{-1}$$

for each unit k in the h -th adjustment group.

The expansion estimator \hat{t}_{EXP} is appropriate for the nonresponse situation: it takes into account the sampling design and it makes an effort to adjust for nonresponse. However, \hat{t}_{EXP} can be improved upon if more information is at hand. Suppose that a single (and always positive) co-variate x is also observed for $k \in s$. In the image of the classical ratio estimator, we can then construct

$$\hat{t}_{\text{RA}} = \left(\sum_s \frac{x_k}{\pi_k} \right) \frac{\sum_{h=1}^H f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}}{\sum_{h=1}^H f_h^{-1} \sum_{r_h} \frac{x_k}{\pi_k}} = \hat{N} \tilde{x}_s \frac{\tilde{y}_r}{\tilde{x}_r},$$

say, where the tilde mean \tilde{x}_r is formed according to (2.1) with x_k instead of y_k , and

$$\tilde{x}_s = \frac{\sum_s \frac{x_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}}.$$

The tilde mean \tilde{x}_s , being formed at the level of the intended sample s , employs sample weights only. (This type of mean can be calculated for the x -variable, which is observed for all $k \in s$, but obviously not for y -variable, which is observed for $k \in r$ only.)

The classical regression estimator formula corresponds, in our context, to

$$\hat{t}_{\text{REG}} = \hat{N} \{ \tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r) \}$$

with

$$b = \frac{\sum_{h=1}^H f_h^{-1} \sum_{r_h} (y_k - \tilde{y}_r) (x_k - \tilde{x}_r) / \pi_k}{\sum_{h=1}^H f_h^{-1} \sum_{r_h} (x_k - \tilde{x}_r)^2 / \pi_k}.$$

(Note: sample weighting as well as nonresponse weighting is used in b too.)

In summary, we have a series of three estimators

$$\hat{t}_{\text{EXP}} = \hat{N} \bar{y}_r, \quad (2.2a)$$

$$\hat{t}_{\text{RA}} = \hat{N} \bar{x}_s \frac{\bar{y}_r}{\bar{x}_r}, \quad (2.2b)$$

$$\hat{t}_{\text{REG}} = \hat{N} \{ \bar{y}_r + b (\bar{x}_s - \bar{x}_r) \}. \quad (2.2c)$$

All three are properly sample weighted and nonresponse weighted. The obvious differences have to do with the co-variate: \hat{t}_{EXP} uses no co-variate, whereas \hat{t}_{RA} and \hat{t}_{REG} do. It is also clear that \hat{t}_{RA} appeals to an underlying relationship between y and the co-variate x in the form of a line through the origin, the slope of which is estimated by \bar{y}_r/\bar{x}_r . In the case of \hat{t}_{REG} , the relationship is a regression with a non-zero intercept. We shall further explore the role of the co-variate.

If the population size N is known, it is in general better to replace \hat{N} by N in (2.2a) to (2.2c), yielding

$$\hat{t}_{\text{EXP}}^* = N \bar{y}_r, \quad (2.3a)$$

$$\hat{t}_{\text{RA}}^* = N \bar{x}_s \frac{\bar{y}_r}{\bar{x}_r}, \quad (2.3b)$$

$$\hat{t}_{\text{REG}}^* = N \{ \bar{y}_r + b (\bar{x}_s - \bar{x}_r) \}. \quad (2.3c)$$

For estimating the population total, N must be known in these three estimators, which may not be the case. However, for estimating the population mean \bar{Y} , they lead, by dividing by N , to the convenient expressions

$$\hat{\bar{Y}}_{\text{EXP}} = \bar{y}_r, \quad (2.4a)$$

$$\hat{\bar{Y}}_{\text{RA}} = \bar{x}_s \frac{\bar{y}_r}{\bar{x}_r}, \quad (2.4b)$$

$$\hat{\bar{Y}}_{\text{REG}} = \bar{y}_r + b (\bar{x}_s - \bar{x}_r). \quad (2.4c)$$

The three series of estimators (2.2), (2.3), and (2.4) are easy to accept on intuitive grounds since all that is involved are elementary weighting principles, plus standard ratio feature or regression feature. Somewhat less elementary is to draw the proper consequences for variance estimation and the construction of valid confidence intervals. These questions are discussed in Section 4. (Contrary to what the rather informal presentation of the estimators (2.2) to (2.4) may suggest, the formulas are not "ad hoc" but the result of a formalized general estimation procedure (with a multivariate regression) for two phases of selection; see Särndal and Swensson (1985a). Most importantly, the variance estimators and confidence intervals follow directly from this theory.)

3. RESPONSE MODELS

The nonresponse weights in the estimators seen in Section 2 can be justified through a response mechanism model involving individual response probabilities that are constant for each unit in a given group. More formally, consider the response mechanism:

MODEL A:

- (1) The probability of response is constant (and equal to an unknown constant Θ_h) for all units $k \in s_h$; $h = 1, \dots, H$.
- (2) The units respond independently of each other.

The theoretical response probabilities Θ_h may vary considerably between groups. (An indication that large differences in response propensity may exist between different subsets is, of course, an incentive to set up adjustment groups, and to weight accordingly.)

Consider a fixed sample realization, s . The group frequencies $n_1, \dots, n_h, \dots, n_H$ are then fixed. Let us also consider a fixed value of the vector of group response frequencies $\underline{m} = (m_1, \dots, m_h, \dots, m_H)$. With s and \underline{m} fixed, the "selection" under Model A of a response set r_h can be shown to conform to a simple random selection of m_h from n_h . The conditional response probability of a unit k in the h -th group is therefore

$$\pi_{k|s,\underline{m}} = \frac{m_h}{n_h} = f_h, \text{ all } k \in s_h. \quad (3.1)$$

(This consideration underlies the weight f_h^{-1} used in the estimators.) Similarly one can show that given s and \underline{m} , the probability under Model A that units k and l respond is

$$\pi_{kl|s,\underline{m}} = \begin{cases} f_h & \text{if } k = l \\ \frac{f_h(m_h - 1)}{n_h - 1} & \text{if } k \neq l \in s_h \\ f_h f_{h'} & \text{if } k \in s_h; l \in s_{h'} \text{ (} h \neq h' \text{)} \end{cases} \quad (3.2)$$

($\pi_{kk|s,\underline{m}}$ is by definition equal to $\pi_{k|s,\underline{m}}$.) These quantities (which remind us of stratified random sampling with m_h units chosen from n_h in the h -th stratum) are important for the calculation of variance estimates and standard errors; see below.

In practice, the analyst decides how to set up his groups s_h . The decision is crucial, for it will determine the adjustment weights f_h^{-1} , and thus the numerical value of the estimate of t , the variance estimate, and the confidence interval. Two different groupings may lead to widely different point estimates and confidence intervals.

The analyst is not so naive as to think that response probabilities exist that are exactly equal within the group that he has identified. He does, however, believe (and usually with good reason) that more valid point estimates and confidence intervals will result with these groups (and thereby the weights f_h^{-1}) than without them. The adjustment group approach is a sound and firmly established practice.

On closer scrutiny, several things may be wrong with a response model such as Model A: the response probability is perhaps not constant within groups. And, even if it were, the particular groups postulated by the model are perhaps wrongly defined; there should have been more groups than assumed, etc. Two cases must therefore be distinguished for the continued discussion:

- (a) The assumed response mechanism (ARM; here in the form of Model A) is true. In practice, this is unlikely to be exactly the case.
- (b) The ARM is more or less false. This is the unpleasant truth in the majority of all practical situations, and it leads to nonresponse bias. In the case of Model A, the groups may be formed more or less incorrectly.

As is usual in statistics, the statistical analyst will formulate the model corresponding to the best of his judgement; accordingly, he will draw certain inferences (confidence statements, for example). Then he will wonder about the robustness of these conclusions, that is, how well do they hold up if the model is false? In the same order of things, let us consider these questions in our particular situation.

4. VARIANCE ESTIMATORS BASED ON A CERTAIN ASSUMED RESPONSE MECHANISM

Model A, with a specified set of groups, is assumed to hold. The response rates, $f_h = m_h/n_h$, $h = 1, \dots, H$, have been established. With this as a starting point, let us examine the variance estimators needed to construct a confidence interval at a specified $100(1 - \alpha)\%$ level. If \hat{t} is one of the estimators in Section 2, and Model A really holds, we have:

- (a) \hat{t} is unbiased (except for a usually unimportant technical bias)
- (b) an approximately $100(1 - \alpha)\%$ confidence interval for t is:

$$\hat{t} \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{t})},$$

where the constant $z_{1-\alpha/2}$ is exceeded with probability $\alpha/2$ by the unit normal variate.

Under repeated draws of samples s and, for each fixed s , repeated realizations (obeying the assumed Model A) of response sets r , the interval will contain the true population total $100(1 - \alpha)\%$ of the time.

The variance and the estimated variance will be determined by two sets of selection probabilities:

1. π_k and π_{kl} , the probabilities of inclusion (first and second order) that accompany the sampling phase;
2. $\pi_{k|s, \underline{m}}$, $\pi_{kl|s, \underline{m}}$ the conditional response probabilities (first and second order) associated with the response Model A ("the nonresponse phase").

In our case, as a consequence of Model A, $\pi_{k|s, \underline{m}}$ and $\pi_{kl|s, \underline{m}}$ are given, respectively, by (3.1) and (3.2). As for π_k and π_{kl} , full generality is assumed; any design may be used for the sampling phase.

A detailed analysis will show that the total variance of any one of the estimators \hat{t} seen in Section 2 can be broken down into two components:

$$V(\hat{t}) = V_1(\hat{t}) + V_2(\hat{t})$$

where $V_1(\hat{t})$ may be termed the sampling variance and $V_2(\hat{t})$ the nonresponse variance. The exact formulas given in Särndal and Swensson (1985a) are not reproduced here, but one notes that the components have some reasonable properties:

1. $V_1(\hat{t}) = 0$ if the whole population U is observed (a census rather than a sample survey);

2. $V_2(\hat{t}) = 0$ if the response is complete ($r = s$);
3. $V_2(\hat{t})$ is greatly reduced in the presence of a strong co-variate, but $V_1(\hat{t})$ is not affected by the co-variate (naturally enough, since it is observed for $k \in s$ only).

Let us examine somewhat more closely the variance estimators. If $\hat{V}_i(\hat{t})$ denotes the estimator of $V_i(\hat{t})$, $i = 1, 2$, the total variance $V(\hat{t})$ will be estimated by an expression of the form

$$\hat{V}(\hat{t}) = \hat{V}_1(\hat{t}) + \hat{V}_2(\hat{t}).$$

Here, the estimated sampling variance component is

$$\hat{V}_1(\hat{t}) = \sum_{k \in r} \sum_{l \in r} \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}} \right) \frac{1}{\pi_{kl|s,m}} u_k u_l,$$

where $\pi_{kl|s,m}$ is given by (3.2), and π_k , π_{kl} are the inclusion probabilities of the sampling design. The estimated nonresponse variance component is

$$\hat{V}_2(\hat{t}) = \sum_{h=1}^H n_h^2 \left(\frac{1}{m_h} - \frac{1}{n_h} \right) S_{wrh}^2$$

with

$$S_{wrh}^2 = \frac{1}{m_h - 1} \sum_{r_h} (w_k - \bar{w}_{r_h})^2$$

The quantities u_k and w_k differ from one estimator \hat{t} to another. Let us look first at the estimated nonresponse variance, $\hat{V}_2(\hat{t})$. This component is of "stratified form": the factor $n_h^2(1/m_h - 1/n_h)$ is characteristic of a stratified simple random selection with m_h units chosen from n_h in the h -th stratum. The reason for this structure lies in the conditional response probabilities $\pi_{kl|s,m}$ given by (3.2).

The quantities w_h have the following appearance:

$$\text{For } \hat{t}_{\text{EXP}} \text{ and } \hat{t}_{\text{EXP}}^*: w_k = \frac{y_k - \bar{y}_r}{\pi_k},$$

$$\text{For } \hat{t}_{\text{RA}} \text{ and } \hat{t}_{\text{RA}}^*: w_k = \frac{y_k - (\bar{y}_r / \bar{x}_r) x_k}{\pi_k},$$

$$\text{For } \hat{t}_{\text{REG}} \text{ and } \hat{t}_{\text{REG}}^*: w_k = \frac{y_k - \bar{y}_r - b(x_k - \bar{x}_r)}{\pi_k}.$$

The expressions for w_k are sample weighted regression residuals. Consequently, if x_k is a powerful explanatory variable for y_k , one will ordinarily have that the variance of the w_k (and thus $\hat{V}_2(\hat{t})$) is smaller for the RA and REG estimators than for the EXP estimator, where the quantity w_k is just a deviation of y_k from the response set mean \bar{y}_r . Consequently, in fortunate circumstances, the part of the standard error that is due to the nonresponse will be reduced to near-zero levels, namely, when x and y have near perfect correlation.

The estimated sampling variance component $\hat{V}_1(\hat{t})$ is of less interest in this discussion, since it is not directly influenced by the co-variate. It should be mentioned, however, that the

u_k are determined as follows: \hat{t}_{EXP} , \hat{t}_{RA} , and \hat{t}_{REG} , $u_k = y_k$, while for the “starred” series of estimators \hat{t}_{EXP}^* , \hat{t}_{RA}^* , and \hat{t}_{REG}^* , $u_k = y_k - \hat{y}_s$, where $\hat{y}_s = (\sum_s \hat{y}_k / \pi_k) / (\sum_s 1 / \pi_k)$ is the mean of the predicted values from the regression fit, so that for \hat{t}_{EXP}^* , $\hat{y}_k = \hat{y}_r$ for all k ; for \hat{t}_{RA}^* , $\hat{y}_k = (\hat{y}_r / \tilde{x}_r) x_k$; and for \hat{t}_{REG}^* , $\hat{y}_k = \hat{y}_r - b(x_k - \tilde{x}_r)$.

A special case arises when $m_h = n_h$ for all h (that is, no nonresponse). Then $\hat{V}_2(\hat{t}) = 0$ (as is reasonable), and $\pi_{kl|s,m} = 1$ for all k and l , leaving the non-zero component

$$\hat{V}_1(\hat{t}) = \sum_{k \in r} \sum_{l \in r} \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}} \right) u_k u_l$$

which is the well-known variance estimator for the case of full response.

5. ROBUSTNESS PROPERTIES WHEN THE ASSUMED RESPONSE MECHANISM IS FALSE

Unbiased estimates and valid confidence intervals can be obtained with the aforementioned estimators, provided the ARM (given by Model A) holds. The presence of a strong co-variate brings about a reduction of the nonresponse component of the variance.

More interesting in a real-life situation is the case where the ARM breaks down. This case must be considered, because even the most careful judgement in setting up adjustment groups is bound to be less than perfect. The extent of the departure of the true response behaviour from that of the ARM will now determine behaviour of the various estimators. The statistical properties (bias, coverage rate achieved by confidence intervals, etc.) are in other words functions of the extent of model breakdown.

In Särndal and Swensson (1985a), a small scale Monte Carlo experiment was carried out to study the impact of certain types of breakdown in Model A. For purposes of illustration, we cite a few results from this study.

The true ARM in the experiment had $H = 4$ adjustment groups, with different response probabilities between groups (but constant response probability for all units in the same group). 1,000 simple random samples were drawn, and each sample was exposed to simulated nonresponse according to the true ARM (which is taken as known, since this is a controlled experiment).

As expected from theory, when the ARM underlying \hat{t}_{EXP} and \hat{t}_{RA} was true, there is essentially no bias, and the empirical coverage rates of the confidence intervals agree essentially with the nominal 95% rate. The advantage of \hat{t}_{RA} lies in a smaller component of variance due to nonresponse. (See “ARM is true” in Table 1.)

False ARM’s were created by joining together groups of the true ARM. The estimator and the confidence interval (based on the false ARM) will then be calculated on the basis of fewer groups than ought to be the case. The case “ARM is false” in Table 1 represents the extreme situation where all four groups of the true ARM were joined into one, meaning that one acts in the estimation process as if all units throughout the population had the same (unknown, but estimated) response probability. The table shows that the co-variate estimator, \hat{t}_{RA} , when compared to the no-co-variate estimator, \hat{t}_{EXP} , has the following (not unexpected) advantages: (a) strong resistance to nonresponse bias (1.26 versus 4.85); (b) much better preservation of the nominal 95% confidence coefficient (92.6% versus 46.3% empirical coverage rate). In addition, \hat{t}_{RA} has a variance advantage, and therefore shorter confidence intervals on the average.

Table 1
Comparison of \hat{t}_{EXP} and \hat{t}_{RA}

Estimator		Absolute bias	Mean of the variance component \hat{V}_2	Empirical coverage rate (95% nominal)
ARM is true	\hat{t}_{EXP}	0.00	1.99	95.2%
	\hat{t}_{RA}	-0.01	0.78	95.5%
ARM is false	\hat{t}_{EXP}	4.85	2.55	46.3%
	\hat{t}_{RA}	1.26	0.78	92.6%

6. CONCLUSION

In summary, we have argued in this paper that two different categories of variables (observed for k in the intended sample s) are of importance:

- (a) variables suitable for estimating the response mechanism (in the case of Model A, these variables allow the construction of the adjustment groups);
- (b) variables (here called co-variables) that are powerful predictors of the y -variable; when used in the estimator formula, they reduce variance and improve the robustness properties.

Whenever possible, one should thus be on the outlook for suitable co-variables. One should also note that when several y -totals are to be estimated, the appropriate co-variables may differ from one y -variable to the other, whereas the weighting classes would probably be set up to apply uniformly for all variables of interest.

REFERENCES

LITTLE, R.J.A. (1983). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

SÄRNDAL, C.E., and SWENSSON, B. (1985a). A general view of estimation for two phases of selection. Part I: Randomized subsample selection (Two-phase sampling). Part II: Nonrandomized subsample selection (Nonresponse). Promemorior fran P/STM no. 20, Statistics Sweden.

SÄRNDAL, C.E., and SWENSSON, B. (1985b). Incorporating nonresponse modelling in a general randomization theory approach. *Bulletin of the International Statistical Institute* (45th session), 51:3, 15.2.1-16.

OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit non-response. In *Incomplete Data in Sample Surveys*, Vol. 2, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 143-183.

Ratio Estimation with Subsampling the Nonrespondents

PODURI S.R.S. RAO¹

ABSTRACT

The procedure of subsampling the nonrespondents suggested by Hansen and Hurwitz (1946) is considered. Post-stratification prior to the subsampling is examined. For the mean of a characteristic of interest, ratio estimators suitable for different practical situations are proposed and their merits are examined. Suitable ratio estimators are also suggested for the situations in which the Hard-Core are present.

KEY WORDS: Auxiliary information; Post-stratification; Biases; Mean square errors; Linear model; Hard-Core.

1. INTRODUCTION

Consider a finite population of size N and a random sample of size n drawn without replacement. In surveys on human populations, frequently n_1 units respond on the items under examination, but the remaining $(n - n_1)$ units do not provide any response. The initial survey may be conducted through the mail or telephone calls, perhaps computer-aided.

In Sections 2, 3 and 4, we consider Hansen and Hurwitz's (1946) procedure of subsampling a portion of the $(n - n_1)$ nonrespondents. In this procedure the population is supposed to be consisting of the response stratum of size N_1 and the nonresponse stratum of size $N_2 = (N - N_1)$.

In Section 2, we discuss two procedures for post-stratifying the sampled units, prior to the subsampling of the nonrespondents.

Two ratio estimators for the mean of an item are considered in Section 3. Biases and Mean Square errors of these estimators are compared in Sections 3 and 4. In Section 4, two more ratio estimators, which may be suitable for some practical situations, are proposed and their relative merits are examined.

The Hard-Core problem is considered in Section 5. Six different estimators for this situation are proposed. Optimum conditions suitable for each one of the estimators are briefly described.

2. HANSEN AND HURWITZ'S ESTIMATOR AND POST-STRATIFICATION

Consider a characteristic of interest y_i , $i = (1, 2, \dots, N)$. Let $\bar{Y} = (\sum_1^N y_i)/N$ and $S^2 = \sum_1^N (y_i - \bar{Y})^2 / (N - 1)$ denote the mean and variance of the population. Let $\bar{Y}_1 = (\sum_1^{N_1} y_i) / N_1$ and $S_1^2 = \sum_1^{N_1} (y_i - \bar{Y}_1)^2 / (N_1 - 1)$ denote the mean and variance of the response group. Similarly, let $\bar{Y}_2 = (\sum_1^{N_2} y_i) / N_2$ and $S_2^2 = \sum_1^{N_2} (y_i - \bar{Y}_2)^2 / (N_2 - 1)$ denote the mean and variance of the nonresponse group. The population

¹ P.S.R.S. Rao, Department of Statistics, University of Rochester, Rochester, NY 14627, U.S.A.

mean can be written as $\bar{Y} = W_1 \bar{Y}_1 + W_2 \bar{Y}_2$, where $W_1 = (N_1/N)$ and $W_2 = (N_2/N)$. The sample mean $\bar{y}_1 = (\sum_1^{n_1} y_i)/n_1$ is unbiased for \bar{Y}_1 , but has a bias equal to $W_2(\bar{Y}_1 - \bar{Y}_2)$ in estimating \bar{Y} .

2.1 Subsampling the Nonrespondents

Hansen and Hurwitz (1946) suggest drawing a subsample of size $m = n_2/k$, $k \geq 1$, from the n_2 nonrespondents and assume that responses are available from all of them. The sample mean $\bar{y}_{2m} = (\sum_1^m y_i)/m$ is unbiased for the mean \bar{y}_2 of the n_2 units. The estimator for \bar{Y} suggested by the above authors is

$$\hat{Y}_{HH} = w_1 \bar{y}_1 + w_2 \bar{y}_{2m}, \quad (2.1)$$

where $w_1 = (n_1/n)$ and $w_2 = (n_2/n)$.

For a given set of n_1 respondents and n_2 nonrespondents, this estimator is unbiased for $\bar{y} = w_1 \bar{y}_1 + w_2 \bar{y}_2 = (\sum_1^n y_i)/n$. Thus, it is unbiased for \bar{Y} .

The variance of this estimator is

$$V(\hat{Y}_{HH}) = \frac{(1-f)}{n} S^2 + W_2 \frac{(k-1)}{n} S_{2m}^2, \quad (2.2)$$

where $f = (n/N)$; see Cochran (1977, p. 371).

Let $s_1^2 = \sum_1^{n_1} (y_i - \bar{y}_1)^2 / (n_1 - 1)$ and $s_{2m}^2 = \sum_1^m (y_i - \bar{y}_{2m})^2 / (m - 1)$ denote the variances of the n_1 responses and the m subsampled units. An unbiased estimator of the variance is

$$\begin{aligned} v(\hat{Y}_{HH}) &= \frac{(1-f)}{n} \left[\frac{(n_1 - 1)s_1^2 + (n_2 - k)s_{2m}^2}{n - 1} \right] \\ &+ \frac{(1-f)}{n} \left[\frac{n_1 (\bar{y}_1 - \hat{Y}_{HH})^2 + n_2 (\bar{y}_{2m} - \hat{Y}_{HH})^2}{n - 1} \right] \\ &+ \frac{(N - 1)w_2(k - 1)s_{2m}^2}{N(n - 1)}. \end{aligned} \quad (2.3)$$

This expression can also be obtained from the variance estimators for double sampling and stratification derived by Cochran (1977, p. 333) and Rao (1973); see also Rao (1983).

Post-stratification and subsampling

The $(n - n_1)$ nonrespondents may be classified into $(L - 1)$ strata of sizes (n_2, n_3, \dots, n_L) according to an auxiliary characteristic, or for convenience in sampling at the next phase. Subsamples of size $m_h = (n_h/k_h)$, $k_h \geq 1$, provide the means $\bar{y}_{hm} = \sum_1^{m_h} y_{hi}/m_h$ and variances $s_{hm}^2 = \sum_1^{m_h} (y_{hi} - \bar{y}_{hm})^2 / (m_h - 1)$.

The unbiased estimator for \bar{Y} now is

$$\hat{Y} = \sum_1^L w_h \bar{y}_{hm}, \quad (2.4)$$

where $w_h = (n_h/n)$ and $\bar{y}_{1m} = \bar{y}_1$.

The variance of the above estimator is

$$V(\hat{Y}) = \frac{(1-f)}{n} S^2 + \sum_2^L \frac{W_h(k_h-1)}{n} S_h^2 \quad (2.6)$$

where $S_h^2 = \Sigma_1^{N_h} (y_{hi} - \hat{Y}_h)^2 / (N_h - 1)$. The estimator for the variance is

$$\begin{aligned} v(\hat{Y}) = & \frac{(1-f)}{n} \sum_1^L \frac{(n_h - k_h) s_{hm}^2}{(n-1)} + \frac{(1-f)}{n} \sum_1^L \frac{n_h (\bar{y}_{hm} - \hat{Y})^2}{(n-1)} \\ & + \frac{(N-1)}{N(n-1)} \sum_2^L w_h (k_h - 1) s_{hm}^2, \end{aligned} \quad (2.7)$$

where $k_h = 1$, $y_{1m} = y_1$, and $s_{1m}^2 = s_1^2$ as defined earlier.

Other types of post-stratification may be considered. For instance, the n units, respondents as well as the nonrespondents, may be post-stratified into L strata according to an auxiliary variable. The h -th stratum will now have n_{h1} respondents ($\Sigma_1^L n_{h1} = n_1$) with mean \bar{y}_{h1} and n_{h2} nonrespondents ($\Sigma_1^L n_{h2} = n_2$). A subsample of size $m_{h2} = (n_{h2}/k_h)$ from the n_{h2} units will provide the mean \bar{y}_{h2m} . An unbiased estimator for the mean \bar{Y}_h of the h -th stratum now is

$$\hat{Y}_h = \frac{n_{h1} \bar{y}_{h1} + n_{h2} \bar{y}_{h2m}}{n_h} \quad (2.8)$$

where $n_h = (n_{h1} + n_{h2})$, and the unbiased estimator for \bar{Y} is

$$\hat{Y} = \sum_1^L \frac{n_h}{n} \hat{Y}_h = \sum_1^L \frac{n_{h1} \bar{y}_{h1} + n_{h2} \bar{y}_{h2m}}{n} \quad (2.9)$$

The variance of this estimator and its estimate can be found as in the above case.

The estimator in (2.4) is preferable if there is much difference among the means of the response and nonresponse strata. The estimator in (2.9) should be preferred if the means of the respondents and nonrespondents differ in each stratum, and if there is much difference among the means of the strata.

Sarndal and Swensson (1985) consider unequal probabilities of selection at the first phase and subsampling the nonrespondents after post-stratification.

3. RATIO ESTIMATORS

Let x_i , $i = (1, 2, \dots, N)$, denote an auxiliary characteristic with population mean $\bar{X} = (\Sigma_1^N x_i) / N$. Let \bar{X}_1 and \bar{X}_2 denote the means of the response and nonresponse groups. Let $\bar{x} = (\Sigma_1^n x_i) / n$ denote the mean of all the n units. Let $\bar{x}_1 = (\Sigma_1^{n_1} x_i) / n_1$ and $\bar{x}_2 = (\Sigma_1^{n_2} x_i) / n_2$ denote the means of the n_1 responding units and the n_2 nonresponding units. Further, let $\bar{x}_{2m} = (\Sigma_1^{m_2} x_i) / m$ denote the mean of the $m = (n_2/k)$ subsampled units.

The population variances of x and y are denoted by S_x^2 and S_y^2 , and the population covariance by S_{xy} . The correlation coefficient is $\rho_{xy} = (S_{xy}/S_x S_y)$. The sample variances are denoted by s_x^2 and s_y^2 . As before, the subscripts 1 and 2 denote the response and nonresponse groups.

3.1 The Conventional Estimator for the Mean

The ratio estimator for \bar{Y} is

$$t_1 = \frac{\bar{y}^*}{\bar{x}^*} \bar{X} = r^* \bar{X} \quad (3.1)$$

where \bar{y}^* is the same as \bar{Y}_{HH} in (2.1), $\bar{x}^* = (w_1 \bar{x}_1 + w_2 \bar{x}_{2m})$, and $r^* = (\bar{y}^*/\bar{x}^*)$; see Cochran (1977, p. 374). Now,

$$t_1 - \bar{Y} = \frac{(\bar{y}^* - R\bar{x}^*) \bar{X}}{\bar{x}^*} \doteq (\bar{y}^* - R\bar{x}^*) \left(1 - \frac{\bar{x}^* - \bar{X}}{\bar{X}}\right) \quad (3.2)$$

where $R = (\bar{Y}/\bar{X})$. The approximation in (3.2) is obtained by expressing $(1/\bar{x}^*)$ in Taylor's series, and it is valid for large values of the sample sizes n and m . From (3.2) the bias of t_1 is

$$B_1 = E(t_1 - \bar{Y}) \doteq \frac{(1-f)}{n\bar{X}} (RS_x^2 - S_{xy}) + \frac{W_2(k-1)}{n\bar{X}} (RS_{x2}^2 - S_{xy2}). \quad (3.3)$$

The bias vanishes only if (a) the regression of y on x goes through the origin for both the response and nonresponse strata and (b) the slopes of both the regressions are equal to R . The first condition is needed for the ratio estimator to be the optimum estimator for \bar{Y} . For the second condition to be satisfied, $R_2 = (\bar{Y}_2/\bar{X}_2)$ should not differ much from $R_1 = (\bar{Y}_1/\bar{X}_1)$.

From (3.2), a large sample approximation to the Mean Square Error (MSE) of t_1 is

$$M_1 = E(t_1 - \bar{Y})^2 \doteq \frac{(1-f)}{n} S_d^2 + W_2 \frac{(k-1)}{n} S_{d2}^2 \quad (3.4)$$

$$= \frac{(1-f)}{n} \sum_1^2 \frac{(NW_h - 1)}{(N-1)} S_{dh}^2 + W_2 \frac{(k-1)}{n} S_{d2}^2 \quad (3.4a)$$

where $S_d^2 = \Sigma_1^N (y_i - Rx_i)^2 / (N-1)$ and $S_{dh}^2 = \Sigma_1^{N_h} (y_{hi} - Rx_{hi})^2 / (N_h - 1)$ for $h = 1, 2$. The expression in (3.4) is briefly indicated by Cochran (1977).

An estimator for this MSE is obtained by replacing S_d^2 in (3.4a) by $s_{d1}^2 = \Sigma_1^{n_1} (y_i - r^* x_i)^2 / (n_1 - 1)$, S_{d2}^2 by $s_{d2}^2 = \Sigma_1^m (y_i - r^* x_i)^2 / (m - 1)$ and W_h by w_h . It is possible to suggest alternative estimators for the above MSE.

3.2 An Alternative Estimator for the Mean

In some situations, there may not be any nonresponse on the auxiliary characteristic. Family size, years of education, years of employment, and the like, are the above type of auxiliary variables.

The subsample provides the means \bar{x}_{2m} and \bar{y}_{2m} . However, since $\bar{x} = (\sum_1^n x_i)/n$ is available, for \bar{Y} we may consider

$$t_2 = \frac{\bar{y}^*}{\bar{x}} \bar{X} = \frac{w_1 \bar{y}_1 + w_2 \bar{y}_{2m}}{\bar{x}} \bar{X}. \quad (3.5)$$

Since the expectation of \bar{y}^* conditional on the first sample is equal to \bar{y} , the bias in t_2 is the same as the one in $\hat{Y}_R = (\bar{y}/\bar{x}) \bar{X}$. We note that \hat{Y}_R is the ratio estimator for the case of complete response. This result can also be derived from the expression

$$t_2 - \bar{Y} = \frac{\bar{y} - R\bar{x}}{\bar{x}} \bar{X} + \frac{\bar{y}^* - \bar{y}}{\bar{x}} \bar{X}. \quad (3.6)$$

Since the conditional mean of \bar{y}^* is equal to \bar{y} , the bias of t_2 is

$$B_2 = E(t_2 - \bar{Y}) = \frac{(1-f)}{n\bar{X}} (RS_x^2 - S_{xy}). \quad (3.7)$$

If the regression of y on x for the entire population goes through the origin, the bias of t_2 in (3.7) vanishes. If the regression for the second stratum also goes through the origin, the bias of t_1 in (3.3) would be small only when $R_2 = (\bar{Y}_2/\bar{X}_2)$ is close to R .

From (3.6), the MSE of t_2 is

$$M_2 = E(t_2 - \bar{Y})^2 = \frac{(1-f)}{n} S_d^2 + \frac{W_2(k-1)}{n} S_{y2}^2 \quad (3.8)$$

$$= \frac{(1-f)}{n} \frac{\sum (NW_h - 1) S_{dh}^2}{N-1} + W_2 \frac{(k-1)}{n} S_{y2}^2. \quad (3.8a)$$

Note that $S_d^2 = S_y^2 + R^2 S_x^2 - 2RS_{xy}$. An estimator of this MSE is obtained by replacing S_{d1}^2 , S_{d2}^2 , S_{y2}^2 , and W_h by s_{d1}^2 , s_{d2}^2 , s_{y2}^2 , and w_h respectively, where

$$s_{d1}^2 = \sum_1^{n_1} (y_i - r^{**} x_i)^2 / (n_1 - 1),$$

$$s_{d2}^2 = \sum_1^m (y_i - r^{**} x_i)^2 / (m - 1),$$

$$s_{y2}^2 = \sum_1^m (y_i - \bar{y}_{2m})^2 / (m - 1).$$

In these expressions, $r^{**} = (\bar{y}^*/\bar{x})$.

Comparing the approximate expressions in (3.4) and (3.8), we find that when $R_1 = (\bar{Y}_1/\bar{X}_1)$ does not differ much from $R_2 = (\bar{Y}_2/\bar{X}_2)$, t_2 will have smaller MSE than t_1 provided the correlation ρ_2 in the nonresponse stratum is not too high. Secondly, if R_1 differs much from R_2 , t_2 may have smaller MSE than t_1 even when ρ_2 is high. The following Section contains further comparisons between these two estimators.

3.3 Further Comparisons

In this Section, we compare t_1 and t_2 through the linear model. For the two groups, we consider the models

$$y_{1i} = \alpha_1 + \beta x_i + e_{1i}, \quad i = (1, 2, \dots, N_1) \tag{3.9a}$$

and

$$y_{2i} = \alpha_2 + \beta x_i + e_{2i}, \quad i = (1, 2, \dots, N_2), \tag{3.9b}$$

with the following assumptions:

$$E(e_{1i} \mid x_i) = 0, \quad E(e_{1i} e_{1i'}) = 0, \quad V(e_{1i} \mid x_i) = v_1 x_i^\ell;$$

$$E(e_{2i} \mid x_i) = 0, \quad E(e_{2i} e_{2i'}) = 0, \quad V(e_{2i} \mid x_i) = v_2 x_i^\ell.$$

We note that $(i \neq i')$ and in practice ℓ may lie between zero and 2. Further e_{1i} and e_{2i} are assumed to be uncorrelated. Biases and MSE's of t_1 and t_2 are obtained in the Appendix with the assumption that the response group of size N_1 and the nonresponse group of size N_2 are samples from the super-populations represented by the above models.

Comparisons of the biases

Let I denote the observations from the first initial sample. Since $E[(1/\bar{x}^*) \mid I] \geq (1/\bar{x})$ and $E(1/\bar{x}) \geq (1/\bar{X})$, from (A.2) and (A.3) we find that both t_1 and t_2 overestimate \bar{Y} . Further the bias B_1 of t_1 is larger than the bias B_2 of t_2 . From (A.6) and (A.7),

$$B_1 - B_2 = \frac{\alpha_w W_2 (k - 1) S_{x2}^2}{n \bar{X}^2} \tag{3.10}$$

This difference in the biases increases with the size of the nonresponse stratum and decreases with an increase in the size of the subsample.

Comparison of the MSE's

From (A.9) and (A.20), the difference in the MSE's of t_1 and t_2 is

$$M_1 - M_2 = (A_1 - A_2) - C_2 + (D_1 - D_2). \tag{3.11}$$

From (A.10), (A.21), and (A.22),

$$(A_1 - A_2) - C_2 = [3 V(\alpha_w) + \alpha_w^2 - \beta^2 \bar{X}^2] \frac{W_2 (k - 1)}{n \bar{x}^2} S_{x2}^2. \tag{3.12}$$

We note that

$$\begin{aligned} V(\alpha_w) &= \alpha_1^2 V(w_1) + \alpha_2^2 V(w_2) + 2\alpha_1 \alpha_2 \text{Cov}(w_1, w_2) \\ &= \frac{N - n}{(N - 1)n} (\alpha_1 - \alpha_2)^2 W_1 W_2. \end{aligned} \tag{3.13}$$

The difference in (3.12) becomes large as α_1 and α_2 differ much from each other. A sufficient condition for the right side of (3.12) to be nonnegative is that $\alpha_W > \beta \bar{X}$. Further analysis of this result shows that the above difference becomes large if $C_x = (S_x/\bar{X})$ becomes larger than $C_y = (S_y/\bar{Y})$ as the correlation $\rho_{xy} = (S_{xy}/S_x S_y)$ increases.

From (A.12) and (A.24),

$$\begin{aligned} D_1 - D_2 &= E \{ [2(\delta - \delta^*) + 3(\delta^{*2} - \delta^2)] \bar{e}^{*2} \} \\ &\quad + 2E[\delta^* - \delta - \delta^{*2} + \delta^2] \bar{E} \bar{e}^*. \end{aligned} \quad (3.14)$$

We note that $(\delta^* - \delta) = (\bar{x}^* - \bar{x})/\bar{X} = w_2(\bar{x}_{2m} - \bar{x}_2)/\bar{X}$. Further, $E(\delta^* - \delta) = 0$.

When $\ell = 0$, from (3.14) and the results in (A.14) and (A.17), to $O(n^{-2})$,

$$\begin{aligned} D_1 - D_2 &= 3E[(\delta^{*2} - \delta^2) \bar{e}^{*2}] - 2E[(\delta^{*2} - \delta^2) \bar{E} \bar{e}^*] \\ &= \frac{3W_2(k-1)S_{x2}^2}{n^2\bar{X}^2} (W_1v_1 + kW_2v_2) - \frac{2W_2(k-1)S_{x2}^2}{Nn\bar{X}^2} (W_1v_1 + W_2v_2) \\ &= \{[2(1-f) + 1](W_1v_1 + W_2v_2) + 3(k-1)W_2v_2\} \frac{W_2(k-1)}{n^2\bar{X}^2} S_{x2}^2. \end{aligned} \quad (3.15)$$

This expression clearly is nonnegative.

When $\ell = 1$, from (3.14), (A.15) and (A.16), to $O(n^{-1})$

$$\begin{aligned} D_1 - D_2 &= 2E \left[(\delta - \delta^*) \frac{(w_1v_1\bar{x}_1 + w_2kv_2\bar{x}_{2m})}{n} \right] \\ &\quad + 2E \left[(\delta^* - \delta) \frac{(w_1v_1\bar{x}_1 + w_2v_2\bar{x}_{2m})}{N} \right]. \end{aligned} \quad (3.16)$$

Noting that $E[(\delta^* - \delta) \bar{x}_1|I] = 0$, from (3.16),

$$\begin{aligned} D_1 - D_2 &= -(2/n) E[kw_2^2\bar{x}_{2m}(\bar{x}_{2m} - \bar{x}_2)]v_2 + (2/N) E[w_2^2\bar{x}_{2m}(\bar{x}_{2m} - \bar{x}_2)]v_2 \\ &= -(2/n)kE[w_2^2V(\bar{x}_{2m}|I)]v_2 + (2/N) E[w_2^2V(\bar{x}_{2m}|I)]v_2 \\ &= -\frac{2(Nk-n)W_2(k-1)S_{x2}^2}{Nn^2\bar{X}^2} v_2. \end{aligned} \quad (3.17)$$

Thus, when $\ell = 1$, $D_2 > D_1$. However, the difference in (3.17) becomes negligible when n is large.

The above results suggest that when $\ell = 0$, t_1 has larger MSE than t_2 if α is larger than $\beta\bar{X}$. When $\ell = 1$, t_1 will have larger MSE than t_2 if α is considerably larger than $\beta\bar{X}$.

4. SEPARATE RATIO ESTIMATORS

4.1 The First Estimator

If (\bar{X}_1, \bar{X}_2) are known, the separate ratio estimator for \bar{Y} that can be suggested is

$$\hat{Y}_S = w_1 r_1 \bar{X}_1 + w_2 r_2 \bar{X}_2, \quad (4.1)$$

where $r_1 = (\bar{y}_1 / \bar{x}_1)$ and $r_2 = (\bar{y}_2 / \bar{x}_2)$. However, (\bar{X}_1, \bar{X}_2) can be estimated by (\bar{x}_1, \bar{x}_2) and $(\bar{y}_{2m} / \bar{x}_{2m})$ is an estimator of r_2 . With these estimates, an estimator for \bar{Y} is

$$t_3 = w_1 \bar{y}_1 + w_2 \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2. \quad (4.2)$$

This estimator can be used if \bar{x}_2 is available but \bar{X} is not; however, it does not make use of \bar{x}_1 .

From (4.2)

$$t_3 - \bar{Y} = (\bar{y} - \bar{Y}) + w_2 (\bar{x}_2 / \bar{x}_{2m}) (\bar{y}_{2m} - r_2 \bar{x}_{2m}). \quad (4.3)$$

If m is large, from (4.3) the bias in t_3 is

$$B_3 = E(t_3 - \bar{Y}) = \frac{(k-1)}{n\bar{X}_2} (R_2 S_{x2}^2 - S_{xy2}). \quad (4.4)$$

The MSE of t_3 is

$$M_3 = E(t_3 - \bar{Y})^2 = \frac{(1-f)}{n} S_y^2 + \frac{W_2(k-1)}{n} S_{r2d2}^2 \quad (4.5)$$

where $S_{r2d2}^2 = \Sigma_1^{N_2} (y_i - R_2 x_i)^2 / (N_2 - 1)$.

An estimator for this MSE is obtained by replacing the first term on the right of (4.5) by $v(\bar{y}) = (1-f)s_y^2/n$, S_{r2d2}^2 by $s_{r2d2}^2 = \Sigma_1^m (y_i - r_{2m} x_i)^2 / (m-1)$, where $r_{2m} = (\bar{y}_{2m} / \bar{x}_{2m})$, and W_2 by w_2 .

4.2 The Second Estimator

An estimator that utilizes \bar{X} and \bar{x} is

$$t_4 = t_3 \left(\frac{\bar{X}}{\bar{x}} \right) = \left(w_1 \bar{y}_1 + w_2 \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2 \right) \left(\frac{\bar{X}}{\bar{x}} \right). \quad (4.6)$$

It may be beneficial to consider this estimator since the conditional mean of t_3 for large m is equal to \bar{y} , and hence the conditional expectation of t_4 becomes equal to $(\bar{y} / \bar{x}) \bar{X}$.

From (4.6),

$$t_4 - \bar{Y} = \left(\frac{\bar{y}}{\bar{x}} \bar{X} - \bar{Y} \right) + w_2 \left(\frac{\bar{x}_2}{\bar{x}_{2m}} \right) (\bar{y}_{2m} - r_2 \bar{x}_{2m}) \left(\frac{\bar{X}}{\bar{x}} \right). \quad (4.7)$$

If n and m are large, the bias of t_4 is

$$B_4 = E(t_4 - \bar{Y}) = \frac{(1-f)}{n\bar{X}} (RS_x^2 - S_{xy}) + \frac{(k-1)}{n\bar{X}_2} (R_2 S_{x2}^2 - S_{xy2}). \quad (4.8)$$

The MSE of t_4 is

$$\begin{aligned} M_4 &= E(t_4 - \bar{Y})^2 = \frac{(1-f)}{n} S_d^2 + W_2 \frac{(k-1)}{n} S_{r2d2}^2 \\ &= \frac{(1-f)}{n} \frac{\sum (NW_h - 1) S_{dh}^2}{N-1} + \frac{W_2 (k-1)}{n} S_{r2d2}^2. \end{aligned} \quad (4.9)$$

An estimator of M_4 is obtained by replacing S_{d1}^2 , S_{d2}^2 , S_{r2d2}^2 , and W_2 by s_{d1}^2 , s_{d2}^2 , s_{r2d2}^2 , and w_2 respectively, where

$$\begin{aligned} s_{d1}^2 &= \sum_{i=1}^{n_1} (y_i - r_{x_i}^*)^2 / (n_1 - 1), \\ s_{d2}^2 &= \sum_{i=1}^m (y_i - r_{x_i}^*)^2 / (m - 1), \\ s_{r2d2}^2 &= \sum_{i=1}^m (y_i - r_{2m} x_i)^2 / (m - 1). \end{aligned}$$

We note that $r^* = (\bar{y}^* / \bar{x}^*)$ as defined in Section (3.1).

Comparing (4.5) and (4.9), we find that t_4 will have smaller MSE than t_3 if the population correlation between x and y is high.

Further investigation is needed to evaluate the merits of the above two separate estimators relative to the estimators in the previous Section.

5. RATIO ESTIMATORS IN THE PRESENCE OF THE HARDCORE

It is becoming increasingly apparent that in spite of subsampling the nonrespondents and a number of call-backs, a significant proportion of the sampled units, the hard-core, do not respond to the items in the survey.

For this situation, we consider the population to be composed of three groups of sizes (N_1, N_2, N_3) , $N = \sum_1^3 N_i$, with means $(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)$ and variances $(S_{y1}^2, S_{y2}^2, S_{y3}^2)$. The means and variances for the auxiliary characteristic are $(\bar{X}_1, \bar{X}_2, \bar{X}_3)$ and $(S_{x1}^2, S_{x2}^2, S_{x3}^2)$. The population means of these two items are $\bar{Y} = (W_1 \bar{Y}_1 + W_2 \bar{Y}_2 + W_3 \bar{Y}_3)$ and $\bar{X} = (W_1 \bar{X}_1 + W_2 \bar{X}_2 + W_3 \bar{X}_3)$, where $\sum_1^3 W_i = 1$. Let $R_1 = (\bar{Y}_1 / \bar{X}_1)$, $R_2 = (\bar{Y}_2 / \bar{X}_2)$ and $R_3 = (\bar{Y}_3 / \bar{X}_3)$.

In the initial sample of size n , only n_1 units respond and provide the means (\bar{x}_1, \bar{y}_1) . The number of units (n_2, n_3) in the last two groups are not known, but their sum $(n_2 + n_3) = (n - n_1)$ is known. The means (\bar{x}_2, \bar{x}_3) of the auxiliary characteristic may be known, but (\bar{y}_2, \bar{y}_3) for the item of interest are not observed.

We consider the situation where in the subsample of size $m = (n - n_1) / k$, only m_2 units respond and provide the means $(\bar{x}_{2m}, \bar{y}_{2m})$. The remaining $m_3 = (m - m_2)$ units, the "hard-core", do not respond. Note that m_1 is not defined.

In Rao and Jackson (1984), a number of estimators for \bar{Y} for the above situation are examined, without utilizing the auxiliary information. In this Section, we suggest the following six estimators that utilize the additional information. We briefly present the conditions for which these estimators may be the optimum ones. For the sake of space, we have not presented the derivations for these estimators.

- (I). The difference between R_1 , R_2 and R_3 is negligible. The m_3 units of the third group, the hard-core, is a random subsample of the m_2 respondents at the second phase. In this case,

$$\hat{Y}_{H1} = \frac{n_1 \bar{y}_1 + (n - n_1) \bar{y}_{2m}}{n_1 \bar{x}_1 + (n - n_1) \bar{x}_{2m}} \bar{X}. \quad (5.1)$$

- (II). Same conditions as in I, but poor correlation in the second and third groups. For this case,

$$\hat{Y}_{H2} = \frac{n_1 \bar{y}_1 + (n - n_1) \bar{y}_{2m}}{n \bar{x}} \bar{X}. \quad (5.2)$$

- (III). $\bar{X}_3 = (N_1 \bar{X}_1 + N_2 \bar{X}_2) / (N_1 + N_2)$ and $\bar{Y}_3 = (N_1 \bar{Y}_1 + N_2 \bar{Y}_2) / (N_1 + N_2)$, and (R_1, R_2, R_3) do not differ much from each other. Under these conditions,

$$\hat{Y}_{H3} = \frac{n_1 \bar{y}_1 + k m_2 \bar{y}_{2m}}{n_1 \bar{x}_1 + k m_2 \bar{x}_{2m}} \bar{X}. \quad (5.3)$$

Note that, since $E(m_2/m) = n_2 / (n - n_1)$, an unbiased estimator of n_2 is $[(n - n_1) / m] m_2 = k m_2$.

- (IV). Same conditions as in (III), but poor correlation in the second and third groups. For this case,

$$\hat{Y}_{H4} = \frac{n_1 \bar{y}_1 + k m_2 \bar{y}_{2m}}{(n_1 + k m_2) \bar{x}} \bar{X}. \quad (5.4)$$

- (V). The three ratios differ from one another. The n_3 units of the third group are a random subsample from the n_2 units of the second group. In this case,

$$\hat{Y}_{H5} = \left[\frac{n_1}{n} \bar{y}_1 + \frac{(n - n_1)}{n} \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2 \right] \left(\frac{\bar{X}}{\bar{x}} \right). \quad (5.5)$$

- (VI). The three ratios differ from one another. The n_3 units of the third group are a random subsample from the $(n_1 + n_2)$ units of the first two groups. Under these conditions,

$$\hat{Y}_{H6} = \left(\frac{n_1}{n_1 + k m_2} \bar{y}_1 + \frac{k m_2}{n_1 + k m_2} \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2 \right) \left(\frac{\bar{X}}{\bar{x}} \right). \quad (5.6)$$

While we expect the above conditions to be satisfactory, further research is needed to evaluate the performances of the above six estimators.

ACKNOWLEDGMENTS

The author would like to thank Dr. J.N.K. Rao and Dr. M.P. Singh for their interest in this topic. Thanks also to the referee for making suggestions towards improving the presentation of the material.

APPENDIX: BIASES AND MSE'S UNDER THE SUPER POPULATION MODEL

Let $\alpha_w = W_1\alpha_1 + W_2\alpha_2$, $\alpha_w = w_1\alpha_1 + w_2\alpha_2$,

$$\bar{E} = \sum_1^N e_i/N, \bar{e}_1 = \sum_1^{n_1} e_i/n_1, \bar{e}_{2m} = \sum_1^m e_i/m \text{ and } \bar{e}^* = w_1\bar{e}_1 + w_2\bar{e}_{2m}.$$

Now

$$\bar{Y} = \alpha_w + \beta\bar{X} + \bar{E}, \tag{A.1}$$

$$t_1 - \bar{Y} = \frac{\bar{X}}{\bar{X}^*}\alpha_w - \alpha_w + \frac{\bar{e}^*}{\bar{X}^*}\bar{X} - \bar{E}, \tag{A.2}$$

and

$$t_2 - \bar{Y} = \frac{\bar{X}}{\bar{X}}\alpha_w - \alpha_w + \beta\left(\frac{\bar{X}^*}{\bar{X}} - 1\right)\bar{X} + \frac{\bar{e}^*}{\bar{X}}\bar{X} - \bar{E}. \tag{A.3}$$

1. Biases

Let $\delta^* = (\bar{x}^* - \bar{X})/\bar{X}$ and $\delta = (\bar{x} - \bar{X})/\bar{X}$. Taylor's expansion about \bar{X} gives

$$\frac{\bar{X}}{\bar{x}^*} = 1 - \delta^* + \delta^{*2} \dots \tag{A.4}$$

and

$$\frac{\bar{X}}{\bar{x}} = 1 - \delta + \delta^2 \dots \tag{A.5}$$

With these expansions, from (A.2) and (A.3), to $O(n^{-1})$ the biases of t_1 and t_2 are

$$B_1 = \frac{V(\bar{x}^*)}{\bar{X}^2}\alpha_w = \left[\frac{(1-f)}{n\bar{X}^2}S_x^2 + \frac{W_2(k-1)}{n\bar{X}^2}S_{x2}^2 \right]\alpha_w \tag{A.6}$$

and

$$B_2 = \frac{V(\bar{x})}{\bar{X}^2}\alpha_w = \left[\frac{(1-f)}{n\bar{X}^2}S_x^2 \right]\alpha_w. \tag{A.7}$$

2. Mean Square Error of t_1

From the expansion in (A.4),

$$\left(\frac{\bar{X}}{\bar{x}^*}\right)^2 \doteq 1 - 2\delta^* + 3\delta^{*2}. \tag{A.8}$$

From (A.2), the MSE of t_1 can be written as

$$M_1 = E(t_1 - \bar{Y})^2 = A_1 + D_1, \tag{A.9}$$

where

$$\begin{aligned}
 A_1 &= E\left(\frac{\bar{X}}{\bar{x}}\alpha_w - \alpha_w\right)^2 \\
 &\doteq E\left[(1 - 2\delta^* + 3\delta^{*2})\alpha_w^2\right] + \alpha_w^2 - 2E\left[(1 - \delta^* + \delta^{*2})\alpha_w\right] \\
 &= V(\alpha_w) + [3E(\alpha_w^2) - 2\alpha_w^2][V(\bar{x}^*)/\bar{X}^2] \\
 &= V(\alpha_w) + [3V(\alpha_w) + \alpha_w^2][V(\bar{x}^*)/\bar{X}^2]
 \end{aligned} \tag{A.10}$$

and

$$D_1 = E\left(\frac{\bar{e}^*}{\bar{x}}\bar{X} - \bar{E}\right)^2. \tag{A.11}$$

With the expansions in (A.4) and (A.8)

$$\begin{aligned}
 \left(\frac{\bar{e}^*}{\bar{x}}\bar{X} - \bar{E}\right)^2 &= \left(\frac{\bar{X}}{\bar{x}}\right)^2\bar{e}^{*2} + \bar{E}^2 - 2\left(\frac{\bar{X}}{\bar{x}}\right)\bar{E}\bar{e}^* \\
 &\doteq (1 - 2\delta^* + 3\delta^{*2})\bar{e}^{*2} + \bar{E}^2 - 2(1 - \delta^* + \delta^{*2})\bar{E}\bar{e}^* \\
 &= (\bar{e}^* - \bar{E})^2 - (2\delta^* - 3\delta^{*2})\bar{e}^{*2} + 2(\delta^* - \delta^{*2})\bar{E}\bar{e}^*.
 \end{aligned} \tag{A.12}$$

Now,

$$\bar{e}^{*2} = w_1^2\bar{e}_1^2 + w_2^2\bar{e}_{2m}^2 + 2w_1w_2\bar{e}_1\bar{e}_{2m}. \tag{A.13}$$

Thus, conditional on n_1 and n_2 , when $\ell = 0$,

$$E(\bar{e}^{*2}|n_1, n_2) = \frac{w_1^2}{n_1}v_1 + \frac{kw_2^2}{n_2}v_2 = \frac{w_1v_1 + kw_2v_2}{n}. \tag{A.14}$$

Similarly, when $\ell = 1$,

$$\begin{aligned}
 E(\bar{e}^{*2}|n_1, n_2) &= \frac{w_1^2}{n_1^2}v_1\left(\sum_1^{n_1}x_i\right) + \frac{(kw_2)^2}{n_2^2}v_2\left(\sum_1^mx_i\right) \\
 &= \frac{1}{n}(w_1v_1\bar{x}_1 + w_2kv_2\bar{x}_{2m}).
 \end{aligned} \tag{A.15}$$

Further,

$$\bar{E}\bar{e}^* = \frac{1}{N}\left[\sum_1^{n_1}e_i + \sum_1^me_i + \sum_1^{N-(n_1+m)}e_i\right](w_1\bar{e}_1 + w_2\bar{e}_{2m}) \tag{A.16}$$

From (A.16), when $t = 0$,

$$E(\bar{E}\bar{e}^*) = \frac{1}{N}(w_1 v_1 + w_2 v_2). \quad (\text{A.17})$$

Similarly, when $\ell = 1$,

$$E(\bar{E}\bar{e}^*) = \frac{1}{N}(w_1 v_1 \bar{x}_1 + w_2 v_2 \bar{x}_{2m}). \quad (\text{A.18})$$

3. Mean Square Error of t_2

From the expansion in (A.5)

$$\left(\frac{\bar{X}}{\bar{x}}\right)^2 \doteq 1 - 2\delta + 3\delta^2. \quad (\text{A.19})$$

From (A.10), the MSE of t_2 can be written as

$$M_2 = E(t_2 - \bar{Y})^2 = A_2 + C_2 + D_2. \quad (\text{A.20})$$

With the expansions in (A.5) and (A.19)

$$\begin{aligned} A_2 &= E\left(\frac{\bar{X}}{\bar{x}}\alpha_w - \alpha_w\right)^2 \\ &\doteq E\left[(1 - 2\delta + 3\delta^2)\alpha_w^2\right] + \alpha_w^2 - 2E\left[(1 - \delta + \delta^2)\alpha_w\right]\alpha_w \\ &= V(\alpha_w) + \left[3E(\alpha_w^2) - 2\alpha_w^2\right]\left[V(\bar{x})/\bar{X}^2\right] \\ &= V(\alpha_w) + \left[3V(\alpha_w) + \alpha_w^2\right]\left[V(\bar{x})/\bar{X}^2\right], \end{aligned} \quad (\text{A.21})$$

$$\begin{aligned} C_2 &= \beta^2 E\left(\frac{\bar{x}^* - \bar{x}}{\bar{x}}\right)^2 \bar{X}^2 \\ &\doteq \beta^2 E(\bar{x}^* - \bar{x})^2 = \beta^2 E\left[w_2^2(\bar{x}_{2m} - \bar{x}_2)^2\right] \\ &= \beta^2 W_2 \frac{(k-1)}{n} S_{x2}^2, \end{aligned} \quad (\text{A.22})$$

and

$$D_2 = E\left(\frac{\bar{x}^*}{\bar{x}}\bar{X} - \bar{E}\right)^2. \quad (\text{A.23})$$

With the expansions in (A.5) and (A.19)

$$\left(\frac{\bar{x}^*}{\bar{x}}\bar{e}^* - \bar{E}\right)^2 = (\bar{e}^* - \bar{E})^2 - (2\delta - 3\delta^2)\bar{e}^{*2} + 2(\delta - \delta^2)\bar{E}\bar{e}^*. \quad (\text{A.24})$$

We note that

$$E\left[\frac{\bar{X}}{\bar{x}}(\alpha_w - \alpha_w)\left(\frac{\bar{x}^* - \bar{x}}{\bar{x}}\right)\bar{X}\right] \doteq E[(\bar{x}^* - \bar{x})(\alpha_w - \alpha_w)] = 0. \quad (\text{A.25})$$

REFERENCES

- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.
- HANSEN, M.H., and HURWITZ, W.N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.
- RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- RAO, P.S.R.S. (1983). Randomization approach. In *Incomplete Data in Sample Surveys*, Vol. 2; (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 33-44.
- RAO, P.S.R.S., and JACKSON, J.E. (1984). Estimation through the procedure of subsampling the nonrespondents. Presented at the American Statistical Association Meetings, Philadelphia.
- SÄRNDAL, C.E., and SWENSSON, B. (1985). Incorporating nonresponse modelling in a general randomization theory approach. *Proceedings of the 45th Session of the International Statistical Institute*, Section 15.2.

ACKNOWLEDGEMENTS

The Survey Methodology Journal wishes to thank the following persons who have served as referees during 1986. An asterisk indicates that the person served more than once during the year.

J. Armstrong, <i>Statistics Canada</i>	G. Kalton, <i>University of Michigan</i>
A. Baldwin, <i>Statistics Canada</i>	T.S. Kheoh, <i>University of Western Ontario</i>
D.A. Binder, <i>Statistics Canada</i>	*H. Lee, <i>Statistics Canada</i>
L.S. Cahoon, <i>U.S. Bureau of the Census</i>	G. Lemaître, <i>Statistics Canada</i>
D.W. Chapman, <i>U.S. Bureau of the Census</i>	R.J. Lowe, <i>Statistics Canada</i>
N. Chinnappa, <i>Statistics Canada</i>	R. Platek, <i>Statistics Canada</i>
*H. Choudhry, <i>Statistics Canada</i>	S. Presser, <i>National Science Foundation</i>
S.G. Currie, <i>Statistics Canada</i>	*J.N.K. Rao, <i>Carleton University</i>
J.-P. Dion, <i>University of Quebec</i>	D. Royce, <i>Statistics Canada</i>
P. Foy, <i>Statistics Canada</i>	D.B. Rubin, <i>Harvard University</i>
J. Gambino, <i>Statistics Canada</i>	G. Sande, <i>Statistics Canada</i>
P. Giles, <i>Statistics Canada</i>	C.E. Särndal, <i>University of Montreal</i>
G.J. Goldmann, <i>Statistics Canada</i>	F. Scheuren, <i>Internal Revenue Service</i>
J.G. Graham, <i>Carleton University</i>	A.J. Scott, <i>University of Auckland</i>
G.B. Gray, <i>Statistics Canada</i>	R.H. Shumway, <i>University of California</i>
R. Groves, <i>University of Michigan</i>	K.P. Srinath, <i>Statistics Canada</i>
*M. Hidirolou, <i>Statistics Canada</i>	R. Valliant, <i>U.S. Bureau of Labour Statistics</i>

Acknowledgements are also due to those who assisted during production of the 1986 issues of the Journal. These people include to mention only a few names, J. Dufour, S. Hupé, D. Lemire, P. Létourneau, A. McGuire, P. Pariseau, L. Quinn, P. Tessier.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 10, n° 2) et de noter les points suivants:

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8 1/2 par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp() et log() etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, l).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.
4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

REMERCIEMENTS

La revue *Techniques d'enquête* désire remercier les personnes suivantes, qui ont bien voulu accepter de faire la critique des articles présentés au cours de l'année 1986. Une astérisque indique que la personne a participé plus d'une fois au cours de l'année.

J. Armstrong, <i>Statistique Canada</i>	G. Kalton, <i>University of Michigan</i>
A. Baldwin, <i>Statistique Canada</i>	T.S. Kheoh, <i>University of Western Ontario</i>
D.A. Binder, <i>Statistique Canada</i>	* H. Lee, <i>Statistique Canada</i>
L.S. Cahoon, <i>U.S. Bureau of the Census</i>	G. Lemaître, <i>Statistique Canada</i>
D.W. Chapman, <i>U.S. Bureau of the Census</i>	R.J. Lowe, <i>Statistique Canada</i>
N. Chinappa, <i>Statistique Canada</i>	R. Platek, <i>Statistique Canada</i>
* H. Choudhry, <i>Statistique Canada</i>	S. Presser, <i>National Science Foundation</i>
S.G. Currie, <i>Statistique Canada</i>	* J.N.K. Rao, <i>Carleton University</i>
J.-P. Dion, <i>Université du Québec</i>	D. Royce, <i>Statistique Canada</i>
P. Foy, <i>Statistique Canada</i>	D.B. Rubin, <i>Harvard University</i>
J. Gambino, <i>Statistique Canada</i>	G. Sande, <i>Statistique Canada</i>
P. Giles, <i>Statistique Canada</i>	C.E. Särndal, <i>Université de Montréal</i>
G.J. Goldmann, <i>Statistique Canada</i>	F. Scheuren, <i>Internal Revenue Service</i>
J.G. Graham, <i>Carleton University</i>	A.J. Scott, <i>University of Auckland</i>
G.B. Gray, <i>Statistique Canada</i>	R.H. Shumway, <i>University of California</i>
R. Groves, <i>University of Michigan</i>	K.P. Srinath, <i>Statistique Canada</i>
* M. Hidiroglou, <i>Statistique Canada</i>	R. Valliant, <i>U.S. Bureau of Labour Statistics</i>

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1986. On ne mentionne ici que quelques-uns de ces noms: J. Dufour, S. Hupe, D. Lemire, P. Létourneau, A. McGuire, P. Pariseau, L. Quinn, P. Tessier.

BIBLIOGRAPHIE

- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.
- HANSEN, M.H., et HURWITZ, W.N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.
- RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- RAO, P.S.R.S. (1983). Randomization approach. Dans *Incomplete Data in Sample Surveys*, Vol. 2; (éd. W.G. Madow, I. Olkin, et D.B. Rubin), New York: Academic Press, 33-44.
- RAO, P.S.R.S., et JACKSON, J.E. (1984). Estimation through the procedure of subsampling the nonrespondents. Présenté aux American Statistical Association Meetings, Philadelphia.
- SÄRNDAL, C.E., et SWENSSON, B. (1985). Incorporating nonresponse modelling in a general randomization theory approach. *Proceedings of the 45th Session of the International Statistical Institute*, Section 15.2.

De l'équation précédente, nous déduisons (lorsque $t = 0$),

(A.17)
$$E(\overline{E\varrho^*}) = \frac{1}{l} (w_1 v_1 + w_2 v_2).$$

De même, lorsque $\ell = 1$,

(A.18)
$$E(\overline{E\varrho^*}) = \frac{1}{l} (w_1 v_1 x_1 + w_2 v_2 x_{2m}).$$

3. Erreur quadratique moyenne de t_2

De l'équation (A.5), nous déduisons

(A.19)
$$\left(\frac{X}{x}\right)^2 \equiv 1 - 2\delta + 3\delta^2.$$

Par l'équation (A.10), nous pouvons exprimer l'EQM de t_2 comme suit:

(A.20)
$$M_2 = E(t_2 - Y)^2 = A_2 + C_2 + D_2.$$

Etant donné les équations (A.5) et (A.19), nous pouvons écrire

$$A_2 = E\left(\frac{X}{x}\alpha_w - \alpha_w\right)^2 \\ \equiv E\left[(1 - 2\delta + 3\delta^2)\alpha_w^2 + \alpha_w^2 - 2E\right](1 - \delta + \delta^2)\alpha_w\alpha_w$$

$$= V(\alpha_w) + \left[3E(\alpha_w^2) - 2\alpha_w^2\right]\left[V(x)/X^2\right]$$

(A.21)
$$= V(\alpha_w) + \left[3V(\alpha_w) + \alpha_w^2\right]\left[V(x)/X^2\right],$$

$$C_2 = \beta_2 E\left(\frac{X}{x} - \frac{X}{x^*}\right)^2$$

$$\equiv \beta_2 E(x^* - x)^2 = \beta_2 E\left[w^2(x_{2m} - x_2)^2\right]$$

(A.22)
$$= \beta_2 W_2 \frac{n}{S_2^{x_2}},$$

et

$$D_2 = E\left(\frac{x}{x^*} X - \frac{x}{x^*} \right)^2.$$

Etant donné les équations (A.5) et (A.19), nous pouvons écrire

(A.24)
$$\left(\frac{x}{x^*} \varrho^* - \frac{x}{x^*} \right)^2 = (\varrho^* - \frac{x}{x^*})^2 - (2\delta - 3\delta^2)\varrho^{*2} + 2(\delta - \delta^2)\frac{x}{x^*}\varrho^*.$$

Il est à noter que

(A.25)
$$E\left[\frac{X}{x}(\alpha_w - \alpha_w^w)\left(\frac{x}{x^*} - \frac{x}{x^*}\right)X\right] \equiv E\left[(x^* - x)(\alpha_w - \alpha_w^w)\right] = 0.$$

où

$$A_1 = E \left(\frac{X}{X^*} \alpha_w^* - \alpha_w \right)^2$$

$$= E \left[(1 - 2\delta^* + 3\delta^*2) \alpha_w^2 + \alpha_w^2 - 2E \left[(1 - \delta^* + \delta^*2) \alpha_w \right] \right. \\ \left. = V(\alpha_w) + [3E(\alpha_w^2) - 2\alpha_w^2] [V(X^*)/X^2] \right. \\ \left. = V(\alpha_w) + [3V(\alpha_w) + \alpha_w^2] [V(X^*)/X^2] \right]$$

(A.10)

et

$$D_1 = E \left(\frac{X}{X^*} X - E \right)^2. \tag{A.11}$$

Par les équations (A.4) et (A.8), nous avons

$$\left(\frac{X}{X^*} X - E \right)^2 = \left(\frac{X}{X^*} \right)^2 e^{*2} + E^2 - 2 \left(\frac{X}{X^*} \right) E e^*$$

$$= (1 - 2\delta^* + 3\delta^*2) e^{*2} + E^2 - 2(1 - \delta^* + \delta^*2) E e^*$$

$$= (e^* - E)^2 - (2\delta^* - 3\delta^*2) e^{*2} + 2(\delta^* - \delta^*2) E e^*. \tag{A.12}$$

Or,

$$e^{*2} = w_1^2 e_1^2 + w_2^2 e_{2m}^2 + 2w_1 w_2 e_1 e_{2m}.$$

(A.13)

Ainsi, lorsque $\ell = 0$, étant donné les valeurs de n_1 et de n_2 ,

$$E(e^{*2} | n_1, n_2) = \frac{n_1^2}{w_1^2} v_1 + \frac{n_2^2}{k w_2^2} v_2 = \frac{n}{w_1 v_1 + k w_2 v_2}.$$

(A.14)

De même, lorsque $\ell = 1$,

$$E(e^{*2} | n_1, n_2) = \frac{n_1^2}{w_1^2} v_1 \left(\sum_{i=1}^I x_i \right) + \frac{n_2^2}{(k w_2^2)^2} v_2 \left(\sum_{i=1}^I x_i \right)$$

$$= \frac{1}{I} (w_1 v_1 x_1 + w_2 k v_2 x_{2m}). \tag{A.15}$$

De plus,

$$E e^* = \frac{1}{I} \left[\sum_{i=1}^I e_i + \sum_{i=1}^I e_i + \sum_{i=1}^{N-(n_1+m)} e_i \right] (w_1 e_1 + w_2 e_{2m}) \tag{A.16}$$

ANNEXE: BIAIS ET ERREURS QUADRATIQUES MOYENNES
DANS LE MODELE DE SUPERPOPULATION

Soit $\alpha_w = W_1\alpha_1 + W_2\alpha_2$, $\alpha_w = w_1\alpha_1 + w_2\alpha_2$,

$$\bar{E} = \sum_N e_i/N, \bar{e}_1 = \sum_{n_1}^I e_i/n_1, \bar{e}_{2m} = \sum_m^I e_i/m \text{ et } \bar{e}^* = w_1\bar{e}_1 + w_2\bar{e}_{2m}.$$

Or,

(A.1)

$$Y = \alpha_w + \beta X + E,$$

(A.2)

$$t_1 - Y = \frac{X}{X^*}\alpha_w - \alpha_w + \frac{\bar{e}}{\bar{e}^*}X - E,$$

et

(A.3)

$$t_2 - Y = \frac{X}{X^*}\alpha_w - \alpha_w + \beta\left(\frac{X}{X^*} - 1\right)X + \frac{\bar{e}}{\bar{e}^*}X - E.$$

1. Biais

Soit $\delta^* = (X^* - X)/X$ et $\delta = (x - X)/X$. Le développement en série de Taylor par rapport à X donne

(A.4)

$$\frac{X}{X^*} = 1 - \delta^* + \delta^{*2} \dots$$

(A.5)

$$\frac{x}{X} = 1 - \delta + \delta^2 \dots$$

À l'aide de ces séries et les équations (A.2) et (A.3), nous pouvons définir les biais de t_1 et de t_2 à $O(n^{-1})$

(A.6)

$$B_1 = \frac{V(X^*)}{V(X)}\alpha_w = \left[\frac{(1-f)X^2}{(1-f)X^2} S_x^2 + \frac{nX^2}{W_2(k-1)} S_{x2}^2 \right] \alpha_w$$

et

(A.7)

$$B_2 = \frac{V(X)}{V(X^*)}\alpha_w = \left[\frac{(1-f)X^2}{nX^2} S_x^2 \right] \alpha_w.$$

2. Erreur quadratique moyenne de t_1

À l'aide de l'équation (A.4), nous pouvons écrire

(A.8)

$$\left(\frac{X}{X^*}\right)^2 = 1 - 2\delta^* + 3\delta^{*2}.$$

Avec l'équation (A.2), nous pouvons exprimer l'EQM de t_1 comme suit

(A.9)

$$M_1 = E(t_1 - Y)^2 = A_1 + D_1,$$

(II). Mêmes conditions qu'en (I) mais faible degré de corrélation dans les deuxième et troisième groupes. Dans ce cas,

$$(5.2) \quad \hat{Y}_{H2} = \frac{n\bar{X}}{n_1\bar{Y}_1 + (n - n_1)\bar{Y}_{2m}} X.$$

(III). $\bar{X}_3 = (N_1\bar{X}_1 + N_2\bar{X}_2) / (N_1 + N_2)$ et $\bar{Y}_3 = (N_1\bar{Y}_1 + N_2\bar{Y}_2) / (N_1 + N_2)$, et (R_1, R_2, R_3) différent peu. Dans ces conditions,

$$(5.3) \quad \hat{Y}_{H3} = \frac{n_1\bar{Y}_1 + km_2\bar{Y}_{2m}}{n_1\bar{X}_1 + km_2\bar{X}_{2m}} X.$$

Comme $E(m_2/m) = n_2 / (n - n_1)$, il convient de souligner que $[(n - n_1) / m] m_2 = km_2$ est un estimateur sans biais de n_2 .

(IV). Mêmes conditions qu'en (III) mais faible degré de corrélation dans les deuxième et troisième groupes. Dans ce cas,

$$(5.4) \quad \hat{Y}_{H4} = \frac{n_1\bar{Y}_1 + km_2\bar{Y}_{2m}}{(n_1 + km_2)\bar{X}} X.$$

(V). Les trois ratios diffèrent. Les n_3 unités du troisième groupe forment un sous-échantillon aléatoire des n_2 unités du deuxième groupe. Dans ce cas,

$$(5.5) \quad \hat{Y}_{H5} = \left[\frac{n_1}{n} \bar{Y}_1 + \frac{(n - n_1)}{(n - n_1)\bar{Y}_{2m}} \frac{\bar{X}_{2m}}{\bar{X}_2} \right] \left(\frac{\bar{X}}{\bar{X}_2} \right).$$

(VI). Les trois ratios diffèrent. Les n_3 unités du troisième groupe forment un sous-échantillon aléatoire des $(n_1 + n_2)$ unités des deux premiers groupes. Dans ces conditions,

$$(5.6) \quad \hat{Y}_{H6} = \left(\frac{n_1}{n_1 + km_2} \bar{Y}_1 + \frac{km_2}{n_1 + km_2} \frac{\bar{Y}_{2m}}{\bar{X}_2} \right) \left(\frac{\bar{X}}{\bar{X}_2} \right).$$

Bien que nous croyions que les conditions ci-dessus sont satisfaisantes, il faudra pousser davantage notre recherche pour évaluer le rendement de ces estimateurs.

REMERCIEMENTS

L'auteur tient à exprimer sa gratitude à Dr. J.N.K. Rao et Dr. M.P. Singh de l'intérêt qu'ils ont manifesté pour cette question. Il tient également à remercier l'arbitre qui a fait des suggestions en vue d'une amélioration de la matière.

On obtient un estimateur de M_4 en remplaçant $S_{d1}^2, S_{d2}^2, S_{r2d2}^2$, et W_2 par $s_{d1}^2, s_{d2}^2, s_{r2d2}^2$, et w_2 respectivement, où

$$s_{d1}^2 = \sum_{n_1}^I (y_i - r_{*}^{x_1})^2 / (n_1 - 1),$$
$$s_{d2}^2 = \sum_m^I (y_i - r_{*}^{x_1})^2 / (m - 1),$$
$$s_{r2d2}^2 = \sum_m^I (y_i - r^{x_1} x_i)^2 / (m - 1).$$

Il est à noter que $r^{*} = (y^{*}/x^{*})$ comme nous l'avons vu dans la section (3.1). Si nous comparons les équations (4.5) et (4.9), nous constatons que l'EQM de t_4 sera inférieure à celle de t_3 si le degré de corrélation entre x et y au niveau de la population est élevé. Seules des recherches supplémentaires nous permettront d'analyser les avantages qu'offrent ces deux estimateurs spéciaux par rapport aux estimateurs étudiés dans la section précédente.

5. ESTIMATEURS PAR QUOTIENT POUR LES CAS PROBLÈMES

Malgré le sous-échantillonnage des non-répondants et les visites de rappel, il est de plus en plus évident qu'une proportion appréciable des unités échantillonnées ne répondent pas aux questions de l'enquête (les cas problèmes).

Pour les besoins de la cause, nous définissons une population composée de trois groupes de tailles respectives (N_1, N_2, N_3) , $N = (\sum_{i=1}^3 N_i)$, auxquels correspondent les moyennes $(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)$ et les variances $(S_{y1}^2, S_{y2}^2, S_{y3}^2)$. Les moyennes et les variances relatives à la variable auxiliaire sont $(\bar{X}_1, \bar{X}_2, \bar{X}_3)$ et $(S_{x1}^2, S_{x2}^2, S_{x3}^2)$. Les moyennes correspondantes de la population sont $\bar{Y} = (W_1 \bar{Y}_1 + W_2 \bar{Y}_2 + W_3 \bar{Y}_3)$ et $\bar{X} = (W_1 \bar{X}_1 + W_2 \bar{X}_2 + W_3 \bar{X}_3)$, où $\sum_{i=1}^3 W_i = 1$. Posons $R_1 = (\bar{Y}_1/\bar{X}_1)$, $R_2 = (\bar{Y}_2/\bar{X}_2)$ et $R_3 = (\bar{Y}_3/\bar{X}_3)$.

Dans l'échantillon initial de taille n_1 , seulement n_1 unités répondent au questionnaire; leurs réponses permettent d'établir les moyennes (\bar{x}_1, \bar{y}_1) . On ne connaît pas le nombre d'unités (n_2, n_3) appartenant aux deux autres groupes mais on connaît leur somme $(n_2 + n_3) = (n - n_1)$. Il se peut que l'on connaisse les moyennes (\bar{x}_2, \bar{x}_3) relatives à la variable auxiliaire mais il n'est pas possible d'établir les moyennes (\bar{y}_2, \bar{y}_3) relatives à la caractéristique étudiée.

Prenons le cas où seulement m_2 unités d'un sous-échantillon de taille $m = (n - n_1)/k$, répondent au questionnaire et leurs réponses permettent d'établir les moyennes $(\bar{x}_{2m}, \bar{y}_{2m})$. Le reste du sous-échantillon $m_3 = (m - m_2)$ unités, c'est-à-dire les cas problèmes, ne répondent pas au questionnaire. Remarquez que m_1 n'est pas défini.

Rao et Jackson (1984) analysent un certain nombre d'estimateurs de \bar{Y} pour le même cas mais n'utilisent pas d'information supplémentaire. Dans la présente section, nous proposons six estimateurs qui font intervenir des informations supplémentaires. Nous exposons brièvement les conditions dans lesquelles ces estimateurs peuvent être optimaux. Pour éviter les développements excessifs, nous ne présentons pas les calculs qui ont permis d'obtenir ces estimateurs. (1). La différence entre R_1, R_2 et R_3 est négligeable. Les m_3 unités du troisième groupe, c'est-à-dire les cas problèmes, forment un sous-échantillon aléatoire des m_2 répondants sélectionnés à la deuxième phase. Dans ce cas,

$$\hat{Y}_{HI} = \frac{n_1 \bar{y}_1 + (n - n_1) \bar{y}_{2m}}{n_1 \bar{x}_1 + (n - n_1) \bar{x}_{2m}} \bar{X}. \tag{5.1}$$

où $r_1 = (\bar{y}_1/\bar{x}_1)$ et $r_2 = (\bar{y}_2/\bar{x}_2)$. Or, (\bar{X}_1, \bar{X}_2) peut être estimé par (\bar{x}_1, \bar{x}_2) et $(\bar{y}_{2m}/\bar{x}_{2m})$ est un estimateur de r_2 . Par ces estimations, nous pouvons définir l'estimateur de \bar{Y} suivant:

$$(4.2) \quad t_3 = w_1 \bar{y}_1 + w_2 \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2.$$

Cet estimateur peut être utilisé lorsque \bar{x}_2 est connu mais que \bar{X} ne l'est pas; nous n'avons toutefois pas besoins de x_1 dans le calcul de cet estimateur.

À l'aide de (4.2)

$$(4.3) \quad t_3 - \bar{Y} = (\bar{y} - \bar{Y}) + w_2 (\bar{x}_2/\bar{x}_{2m}) (\bar{y}_{2m} - r_2 \bar{x}_{2m}).$$

Si m est grand, le biais contenu dans t_3 est, d'après l'équation 4.3:

$$(4.4) \quad B_3 = E(t_3 - \bar{Y}) = \frac{(k-1)}{(k-1)} \frac{n \bar{X}_2}{(R_2 S_{x_2}^2 - S_{xy_2})}.$$

L'EQM de t_3 est

$$(4.5) \quad M_3 = E(t_3 - \bar{Y})^2 = \frac{n}{(1-f)} S_y^2 + \frac{n}{W_2(k-1)} S_{x_2}^2$$

où $S_{x_2}^2 = \Sigma (y_i - R_2 x_i)^2 / (N_2 - 1)$.
On obtient un estimateur sans biais de cette EQM en remplaçant le premier terme du membre de droite de l'équation (4.5) par $v(\bar{y}) = (1-f) S_y^2/n$, $S_{x_2}^2$ par $S_{x_2}^2$ et $\Sigma (y_i - R_2 x_i)^2 / (m-1)$, où $r_{2m} = (\bar{y}_{2m}/\bar{x}_{2m})$, et W_2 par w_2 .

4.2 Second estimateur

Nous avons ci-dessous un estimateur qui fait intervenir \bar{X} et \bar{x} :

$$(4.6) \quad t_4 = t_3 \left(\frac{\bar{X}}{\bar{x}} \right) = \left(w_1 \bar{y}_1 + w_2 \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2 \right) \left(\frac{\bar{X}}{\bar{x}} \right).$$

Il peut être utile de considérer cet estimateur puisque la moyenne conditionnelle de t_3 est égale à \bar{y} , lorsque m est grand et que, par conséquent, l'espérance mathématique conditionnelle de t_4 devient égale à $(\bar{y}/\bar{x})\bar{X}$.

Par (4.6),

$$(4.7) \quad t_4 - \bar{Y} = \left(\frac{\bar{Y}}{\bar{x}} \bar{X} - \bar{Y} \right) + w_2 \left(\frac{\bar{x}_2}{\bar{x}_{2m}} \right) (\bar{y}_{2m} - r_2 \bar{x}_{2m}) \left(\frac{\bar{X}}{\bar{x}} \right).$$

Si n et m sont grands, le biais de t_4 est

$$(4.8) \quad B_4 = E(t_4 - \bar{Y}) = \frac{n \bar{X}}{(1-f)} (R S_{x_2}^2 - S_{xy}) + \frac{n \bar{X}_2}{(k-1)} (R_2 S_{x_2}^2 - S_{xy_2}).$$

L'EQM de t_4 est

$$M_4 = E(t_4 - \bar{Y})^2 = \frac{n}{(1-f)} S_y^2 + W_2 \frac{n}{(k-1)} S_{x_2}^2 + \frac{n}{W_2(k-1)} S_{x_2}^2.$$

(4.9)

La différence exprimée en (3.12) augmente à mesure que l'écart entre α_1 et α_2 s'élargit. Pour que le membre de droite de l'équation (3.12) soit non-négatif, il suffit que $\alpha_W > \beta X$. Une analyse approfondie de ce résultat montre que la différence exprimée en (3.12) s'accroît lorsque $C_X = (S_X/X)$ devient supérieur à $C_Y = (S_Y/Y)$ tandis que le coefficient de corrélation $\rho_{XY} = (S_{XY}/S_X S_Y)$ augmente.

Par (A.12) et (A.24),

$$D_1 - D_2 = E\{[2(\delta - \delta^*) + 3(\delta^{*2} - \delta^2)]\bar{e}^{*2}\}$$

$$+ 2E[\delta^* - \delta - \delta^{*2} + \delta^2] \bar{E} \bar{e}^*]. \quad (3.14)$$

Il est à noter que $(\delta^* - \delta) = (X^* - X)/\bar{X} = w_2(\bar{x}_{2m} - \bar{x}_2)/\bar{X}$. De plus, $E(\delta^* - \delta) = 0$.

Lorsque $\ell = 0$, l'écart peut être établi à $O(n^{-2})$ à l'aide des équations (3.14), (A.14) et (A.17),

$$D_1 - D_2 = 3E[(\delta^{*2} - \delta^2)\bar{e}^{*2}] - 2E[(\delta^{*2} - \delta^2)\bar{E} \bar{e}^*]$$

$$= \frac{3W_2(k-1)S_{x_2}^2}{W_1v_1 + kW_2v_2} (W_1v_1 + kW_2v_2) - \frac{NnX_2}{2W_2(k-1)S_{x_2}^2} (W_1v_1 + W_2v_2)$$

$$= \{[2(1-f) + 1](W_1v_1 + W_2v_2) + 3(k-1)W_2v_2\} \frac{W_2(k-1)}{S_{x_2}^2} \frac{n^2X_2}{S_{x_2}^2}.$$

(3.15)

De toute évidence, l'expression ci-dessus est non négative.

Lorsque $\ell = 1$, l'écart à $O(n^{-1})$ peut être établi comme suit à l'aide des équations (3.14), (A.15) et (A.16),

$$D_1 - D_2 = 2E[(\delta - \delta^*) \frac{n}{(w_1v_1\bar{x}_1 + w_2kv_2\bar{x}_{2m})}] + 2E[(\delta^* - \delta) \frac{N}{(w_1v_1\bar{x}_1 + w_2v_2\bar{x}_{2m})}]. \quad (3.16)$$

Étant donné que $E[(\delta^* - \delta)\bar{x}_1|I] = 0$ nous obtenons par l'équation (3.16),

$$D_1 - D_2 = - (2/n)E[kw_2^2\bar{x}_{2m}(\bar{x}_{2m} - \bar{x}_2)]v_2 + (2/N)E[w_2^2\bar{x}_{2m}(\bar{x}_{2m} - \bar{x}_2)]v_2$$

$$= - (2/n)kE[w_2^2V(\bar{x}_{2m}|I)]v_2 + (2/N)E[w_2^2V(\bar{x}_{2m}|I)]v_2$$

$$= - \frac{2(Nk - n)W_2(k-1)S_{x_2}^2}{Nn^2X_2}v_2. \quad (3.17)$$

Ainsi, lorsque $\ell = 1$, $D_2 > D_1$. Toutefois, l'écart devient négligeable lorsque n est grand. Ces résultats nous permettent de supposer que lorsque $\ell = 0$, l'EQM de t_1 est supérieure à celle de t_2 si α est plus grand que βX . Lorsque $\ell = 1$, l'EQM de t_1 sera supérieure à celle de t_2 si α est beaucoup plus grand que βX .

4. ESTIMATEURS PAR QUOTIENT SPÉCIAUX

4.1 Premier estimateur

Si (X_1, X_2) est connu, nous pouvons envisager l'estimateur par quotient spécial de Y suivant:

$$\hat{Y}_S = w_1r_1\bar{X}_1 + w_2r_2\bar{X}_2, \quad (4.1)$$

3.3 Autres comparaisons

Dans cette sous-section, nous comparons t_1 et t_2 à l'aide du modèle linéaire. Pour les deux groupes, nous examinons les modèles

$$y_{1i} = \alpha_1 + \beta x_i + e_{1i}, \quad i = (1, 2, \dots, N_1) \quad (3.9a)$$

et

$$y_{2i} = \alpha_2 + \beta x_i + e_{2i}, \quad i = (1, 2, \dots, N_2), \quad (3.9b)$$

et posons les hypothèses suivantes:

$$\begin{aligned} E(e_{1i}|x_i) &= 0, E(e_{1i}e_{1i'}) = 0, V(e_{1i}|x_i) = v_1x_i^2; \\ E(e_{2i}|x_i) &= 0, E(e_{2i}e_{2i'}) = 0, V(e_{2i}|x_i) = v_2x_i^2. \end{aligned}$$

Il est à noter que $i \neq i'$ et qu'en pratique la valeur de i peut varier de 0 à 2. Nous supposons, de plus, qu'il n'existe aucune corrélation entre e_{1i} et e_{2i} . Nous calculons en annexe les biais et les EQM de t_1 et de t_2 en supposant que le groupe de réponse de taille N_1 et le groupe de non-réponse de taille N_2 sont des échantillons tirés des superpopulations représentées par les modèles ci-dessus.

Comparaison des biais

Soit I les observations de l'échantillon initial. Comme $E[(1/x^*)|I] \geq (1/x)$ et $E(1/x) \geq (1/X)$, nous déduisons de (A.2) et de (A.3) que t_1 et t_2 donnent tous deux une surestimation de Y . En outre, le biais B_1 de t_1 est supérieur au biais B_2 de t_2 . Par (A.6) et

$$B_1 - B_2 = \frac{\alpha_W W_2(k-1)S_{x2}^2}{nX^2} \quad (3.10)$$

Cet écart entre les biais augmente avec la taille de la strate de non-réponse et diminue à mesure que la taille du sous-échantillon augmente.

Comparaison des EQM

À l'aide des équations (A.9) et (A.20), nous pouvons établir la différence entre les EQM de t_1 et de t_2 :

$$M_1 - M_2 = (A_1 - A_2) - C_2 + (D_1 - D_2). \quad (3.11)$$

Par (A.10), (A.21), et (A.22),

$$(A_1 - A_2) - C_2 = [3V(\alpha_W) + \alpha_W^2 - \beta_2 X_2^2] \frac{nX^2}{W_2^2(k-1)S_{x2}^2}. \quad (3.12)$$

Nous constatons que

$$\begin{aligned} V(\alpha_W) &= \alpha_1^2 V(w_1) + \alpha_2^2 V(w_2) + 2\alpha_1\alpha_2 \text{Cov}(w_1, w_2) \\ &= \frac{N}{N-n} \frac{(N-1)n}{(\alpha_1 - \alpha_2)^2 W_1 W_2}. \end{aligned} \quad (3.13)$$

Les moyennes du sous-échantillon seront \bar{x}_{2m} et \bar{y}_{2m} . Toutefois, comme nous connaissons $\bar{x} = (\sum^n_i x_i)/n$, nous pouvons considérer l'estimateur suivant pour \bar{Y} ,

$$t_2 = \frac{\bar{y}}{\bar{y}^*} \bar{X} = \frac{\bar{x}}{w_1 \bar{y}_1 + w_2 \bar{y}_{2m}} \bar{X}. \tag{3.5}$$

Comme l'espérance mathématique de \bar{y}^* liée au premier échantillon est égale à \bar{y} , le biais contenu dans t_2 est identique à celui contenu dans $\hat{Y}_R = (\bar{y}/\bar{x}) \bar{X}$. Nous savons que \hat{Y}_R est l'estimateur par quotient qui s'applique lorsque le taux de non-réponse est nul. Nous arrivons au même résultat lorsque nous utilisons l'expression suivante:

$$t_2 - \bar{Y} = \frac{\bar{y} - R\bar{x}}{\bar{y}^* - \bar{y}} \bar{X} + \frac{\bar{x}}{\bar{y}^*} \bar{X}. \tag{3.6}$$

Comme la moyenne conditionnelle de \bar{y}^* est égale à \bar{y} , le biais contenu dans t_2 est

$$B_2 = E(t_2 - \bar{Y}) = \frac{n\bar{X}}{(1-f)RS^2_x - S_{xy}}. \tag{3.7}$$

Si la droite de régression de y par rapport à x pour l'ensemble de la population passe par l'origine, le biais de t_2 dans l'équation (3.7) disparaît. Si la droite de régression relative à la seconde strate passe aussi par l'origine, le biais de t_1 (équation 3.3) sera faible seulement si $R_2 = (\bar{Y}_2/\bar{X}_2)$ est proche de R .

Par l'équation (3.6), l'EQM de t_2 est

$$M_2 = E(t_2 - \bar{Y})^2 = \frac{(1-f)^2}{n} S^2_{xy} + \frac{S^2_d}{W_2(k-1)} S^2_{y2} \tag{3.8}$$

$$= \frac{(1-f)^2}{n} \frac{\Sigma (NW_h - 1) S^2_{dh}}{N-1} + W_2 \frac{n}{(k-1)} S^2_{y2}. \tag{3.8a}$$

Il faut noter que $S^2_d = S^2_y + R^2 S^2_x - 2RS_{xy}$. On obtient un estimateur de l'EQM ci-dessus en remplaçant S^2_{d1} , S^2_{d2} , S^2_{y2} , et W_h par s^2_{d1} , s^2_{d2} , s^2_{y2} , et w_h respectivement, où

$$s^2_{d1} = \sum_{n_1}^1 (y_i - r^{**}x_i)^2 / (n_1 - 1),$$
$$s^2_{d2} = \sum_m^1 (y_i - r^{**}x_i)^2 / (m - 1),$$
$$s^2_{y2} = \sum_m^1 (y_i - \bar{y}_{2m})^2 / (m - 1).$$

Dans les formules, $r^{**} = (\bar{y}^*/\bar{x})$.

Si nous comparons les formules d'approximation en (3.4) et (3.8), nous constatons que lorsque $R_1 = (\bar{Y}_1/\bar{X}_1)$ se rapproche sensiblement de $R_2 = (\bar{Y}_2/\bar{X}_2)$, l'EQM de t_2 est inférieure à celle de t_1 à condition que le coefficient de corrélation ρ_2 dans la strate de non-réponse ne soit pas trop élevé. En deuxième lieu, si les valeurs de R_1 et de R_2 divergent sensiblement, l'EQM de t_2 pourrait être encore inférieure à celle de t_1 même lorsque ρ_2 est élevé. Dans la sous-section suivante, nous poussons plus loin la comparaison des deux estimateurs.

Les variances de la population de x et de y sont désignées S_x^2 et S_y^2 , et la covariance de la population, S_{xy} . Le coefficient de corrélation est $\rho_{xy} = (S_{xy}/S_x S_y)$. Les variances de l'échantillon sont désignées s_x^2 et s_y^2 . Les indices 1 et 2 désignent toujours respectivement le groupe de réponse et le groupe de non-réponse.

3.1 Estimateur classique de la moyenne

L'estimateur par quotient de \bar{Y} est

$$t_1 = \frac{\bar{y}^*}{X^*} X^* = r^* X^* \quad (3.1)$$

où \bar{y}^* équivaut à \bar{Y}^{HH} défini en (2.1), $X^* = (w_1 \bar{x}_1 + w_2 \bar{x}_{2m})$, et $r^* = (\bar{y}^*/X^*)$; voir Cochran (1977, p. 374). Or,

$$t_1 - \bar{Y} = \frac{(\bar{y}^* - R X^*) X^*}{(\bar{y}^* - R X^*) (1 - R X^*)} = (\bar{y}^* - R X^*) \left(1 - \frac{X^*}{X^* - \bar{Y}}\right) \quad (3.2)$$

où $R = (\bar{Y}/X^*)$. La formule d'approximation ci-dessus est obtenue en développant $(1/X^*)$ en série de Taylor et elle vaut pour des tailles d'échantillon n et m élevées. À l'aide de l'équation (3.2), nous pouvons définir le biais de t_1

$$B_1 = E(t_1 - \bar{Y}) = \frac{(1 - f) n X^*}{W_2 (k - 1)} (R S_x^2 - S_{xy}) + \frac{n X^*}{W_2 (k - 1)} (R S_{x_2}^2 - S_{xy_2}). \quad (3.3)$$

Ce biais disparaît seulement si a) la droite de régression de y par rapport à x passe par l'origine, aussi bien pour la strate de réponse que pour la strate de non-réponse, et que b) les pentes des deux droites de régression sont égales à R . La première condition doit être satisfaite pour que l'estimateur par quotient de \bar{Y} soit optimal. Pour que la deuxième condition soit satisfaisante, la valeur de $R_2 = (\bar{Y}_2/\bar{X}_2)$ ne doit pas être trop différente de celle de $R_1 = (\bar{Y}_1/\bar{X}_1)$.

À l'aide de l'équation (3.2), il est possible d'établir une approximation de l'erreur quadratique moyenne (EQM) de t_1 pour de grands échantillons, notamment:

$$M_1 = E(t_1 - \bar{Y})^2 = \frac{(1 - f) n}{(k - 1)} S_{\bar{y}}^2 + W_2 \frac{n}{(k - 1)} S_{\bar{y}}^2 \quad (3.4)$$

$$= \frac{(1 - f) n}{2} \sum_{i=1}^I \frac{n}{(N W_h - 1)} \frac{N - 1}{S_{\bar{y}}^2} + W_2 \frac{n}{(k - 1)} S_{\bar{y}}^2 \quad (3.4a)$$

où $S_{\bar{y}}^2 = \Sigma_N^1 (y_i - R x_i)^2 / (N - 1)$ et $S_{\bar{y}}^2 = \Sigma_N^{dh} (y_{hi} - R x_{hi})^2 / (N_h - 1)$ pour $h = 1, 2$. Cochran (1977) mentionne brièvement l'expression définie en (3.4).

On obtient un estimateur de l'EQM ci-dessus en remplaçant $S_{\bar{y}}^2$ dans l'équation (3.4a) par $S_{\bar{y}}^2 = \Sigma_N^1 (y_i - r^* x_i)^2 / (n_1 - 1)$, $S_{\bar{y}}^2$ par $S_{\bar{y}}^2 = \Sigma_m^1 (y_i - r^* x_i)^2 / (m - 1)$ et W_h by w_h . D'autres estimateurs de l'EQM peuvent être définis.

3.2 Autre estimateur possible de la moyenne

Il arrive parfois qu'on n'enregistre aucun cas de non-réponse pour une variable auxiliaire. La taille de la famille, le niveau d'instruction, le nombre d'années d'emploi et les autres caractéristiques du même genre sont des exemples de variables auxiliaires pour lesquelles le taux de non-réponse est souvent nul.

La variance de l'estimateur (2.4) est

$$V(\hat{Y}) = \frac{(1-f)}{n} S^2 + \sum_L^2 \frac{W_h(k_h-1)}{n} S_h^2 \quad (2.6)$$

où $S_h^2 = \Sigma_{N_h}^1 (y_{hi} - \bar{Y}_h)^2 / (N_h - 1)$. L'estimateur de la variance est

$$v(\hat{Y}) = \frac{(1-f)}{n} \sum_L^1 \frac{1}{(n_h - k_h) s_{hm}^2} + \frac{(1-f)}{n} \sum_L^1 \frac{1}{n_h (\bar{Y}_{hm} - \bar{Y})^2} + \frac{(N-1)}{N(n-1)} \sum_L^2 w_h(k_h-1) s_{hm}^2 \quad (2.7)$$

où $k_h = 1$, $y_{1m} = y_1$, et $s_{1m}^2 = s_1^2$, comme nous l'avons vu plus tôt.

D'autres modes de stratification a posteriori peuvent être considérés. Par exemple, il est possible de stratifier a posteriori les n unités, (répondants aussi bien que non-répondants) en L strates selon une variable auxiliaire. La h -ième strate contiendra alors n_{h1} répondants ($\Sigma_L^1 n_{h1} = n_1$) avec une moyenne correspondante de \bar{y}_{h1} et n_{h2} non-répondants ($\Sigma_L^1 n_{h2} = n_2$). Un sous-échantillon de taille $m_{h2} = (n_{h2}/k_h)$ prélevé parmi les n_{h2} unités aura pour moyenne \bar{y}_{h2m} . Un estimateur sans biais de la moyenne \bar{Y}_h se rattachant à la h -ième strate est alors défini:

$$\hat{Y}_h = \frac{n_{h1} \bar{y}_{h1} + n_{h2} \bar{y}_{h2m}}{n_h} \quad (2.8)$$

où $n_h = (n_{h1} + n_{h2})$, et l'estimateur sans biais de \bar{Y} est

$$\hat{Y} = \sum_L^1 \frac{n_h \hat{Y}_h}{L} = \sum_L^1 \frac{n}{(n_{h1} \bar{y}_{h1} + n_{h2} \bar{y}_{h2m})} \quad (2.9)$$

La variance de l'estimateur (2.9) et l'estimation de sa valeur peuvent être trouvées en répétant les procédures suivies lors du cas précédent.

Il est préférable d'utiliser l'estimateur défini en (2.4) si les moyennes respectives des strates de réponse et de non-réponse divergent sensiblement. En revanche, il conviendrait d'utiliser l'estimateur défini en (2.9) lorsque les moyennes relatives aux répondants et aux non-répondants diffèrent l'une de l'autre dans chaque strate et qu'il y a un écart prononcé entre les moyennes des strates.

Sarnadal et Swensson (1985) considèrent des probabilités inégales de sélection à la première phase et le sous-échantillonnage des non-répondants après une stratification a posteriori.

3. ESTIMATEURS PAR QUOTIENT

Définissons x_i , $i = (1, 2, \dots, N)$, une variable auxiliaire avec une moyenne de population $\bar{X} = (\Sigma_N^1 x_i) / N$. Définissons X_1 et X_2 les moyennes se rattachant respectivement aux groupes de réponse et de non-réponse. Soit $\bar{x} = (\Sigma_n^1 x_i) / n$ la moyenne des unités et $\bar{x}_1 = (\Sigma_{n_1}^1 x_i) / n_1$ et $\bar{x}_2 = (\Sigma_{n_2}^1 x_i) / n_2$ les moyennes relatives aux n_1 unités répondantes et aux n_2 non-répondantes. En outre, posons $\bar{x}_{2m} = (\Sigma_m^1 x_i) / m$ la moyenne des $m = (n_2/k)$ unités du sous-échantillon.

2.1 Sous-échantillonnage des non-répondants

désignent respectivement la moyenne et la variance du groupe de non-réponse. La moyenne de la population peut être exprimée par $\bar{Y} = W_1 \bar{Y}_1 + W_2 \bar{Y}_2$, où $W_1 = (N_1/N)$ et $W_2 = (N_2/N)$. La moyenne de l'échantillon $\bar{y}_1 = (\sum_{i=1}^n y_i) / n$, est non biaisée dans le cas de \bar{Y}_1 mais comporte un biais égal à $W_2 (\bar{Y}_1 - \bar{Y}_2)$ pour l'estimation de \bar{Y} .

Hansen et Hurwitz (1946) proposent de prélever un sous-échantillon de taille $m = n_2/k$, $k \geq 1$, parmi les n_2 non-répondants et supposent que toutes les unités ainsi échantillonnées acceptent de répondre au questionnaire. La moyenne de l'échantillon $\bar{y}_{2m} = (\sum_{i=1}^m y_i) / m$ est non biaisée pour la moyenne \bar{y}_2 des n_2 unités. L'estimateur de \bar{Y} que les auteurs proposent est

$$\hat{Y}_{HH} = w_1 \bar{y}_1 + w_2 \bar{y}_{2m}, \quad (2.1)$$

où $w_1 = (n_1/n)$ et $w_2 = (n_2/n)$. Pour un ensemble donné de n_1 répondants et de n_2 non-répondants, cet estimateur est non biaisé pour $\bar{Y} = w_1 \bar{Y}_1 + w_2 \bar{Y}_2 = (\sum_{i=1}^n y_i) / n$. Il est, par conséquent, non biaisé pour \bar{Y} . La variance de cet estimateur est

$$V(\hat{Y}_{HH}) = \frac{n}{(1-f)} S_2^2 + W_2 \frac{n}{(k-1)} S_2^2, \quad (2.2)$$

où $f = (n/N)$; voir Cochran (1977, p. 371). Définissons $s_1^2 = \sum_{i=1}^n (y_i - \bar{y}_1)^2 / (n_1 - 1)$ et $s_{2m}^2 = \sum_{i=1}^m (y_i - \bar{y}_{2m})^2 / (m - 1)$ les variances se rattachant respectivement aux n_1 répondants et aux m unités sous-échantillonnées. Un estimateur sans biais de la variance est

$$v(\hat{Y}_{HH}) = \frac{n}{(1-f)} \left[\frac{n-1}{(n_1-1)s_1^2 + (n_2-k)s_{2m}^2} \right] + \frac{n}{(1-f)} \left[\frac{n-1}{n_1(\bar{y}_1 - \hat{Y}_{HH})^2 + n_2(\bar{y}_{2m} - \hat{Y}_{HH})^2} \right] + \frac{(N-1)w_2(k-1)s_{2m}^2}{N(n-1)}.$$

On peut aussi obtenir cette expression à l'aide des estimateurs de la variance pour l'échantillonnage double et la stratification calculés par Cochran (1977, p. 333) et J.N.K. Rao (1973); voir aussi Rao (1983).

Stratification a posteriori et sous-échantillonnage

Les $(n - n_1)$ non-répondants peuvent être répartis en $(L - 1)$ strates de tailles respectives (n_2, n_3, \dots, n_L) selon une variable auxiliaire ou, pour des raisons de commodité, à l'aide d'un échantillonnage effectué à la phase suivante. Les sous-échantillons de taille $m_h = (n_h/k_h)$, $k_h \geq 1$, ont pour moyennes $\bar{y}_{hm} = (\sum_{i=1}^{m_h} y_{hi}) / m_h$ et pour variances $s_{hm}^2 = \sum_{i=1}^{m_h} (y_{hi} - \bar{y}_{hm})^2 / (m_h - 1)$. L'estimateur sans biais de \bar{Y} est désormais défini

$$\bar{Y} = \sum_{h=1}^L w_h \bar{y}_{hm}, \quad (2.4)$$

où $w_h = (n_h/n)$ et $\bar{y}_{1m} = \bar{y}_1$.

Estimation par le quotient dans le cas d'un sous-échantillonnage des non-répondants

PODURI S.R.S. RAO¹

RÉSUMÉ

L'auteur examine le sous-échantillonnage des non-répondants, tel que l'ont proposé Hansen et Hurwitz (1946), et la stratification a posteriori effectuée avant le sous-échantillonnage. Pour l'estimation des moyennes relatives à des caractéristiques étudiées, il propose des estimateurs par quotient adaptés à diverses situations et analyse les avantages de chacun. Des estimateurs par quotient particuliers sont également proposés pour les cas problèmes.

MOTS CLÉS: Information supplémentaire; stratification a posteriori; biais; erreur quadratique moyenne; modèle linéaire; cas problème.

1. INTRODUCTION

Considérons une population finie de taille N et un échantillon aléatoire de taille n prélevé sans remise. Dans les enquêtes démographiques, il arrive souvent que n_1 unités répondent au questionnaire d'enquête mais que le reste des unités choisies, soit $(n - n_1)$, n'y répondent pas. L'enquête initiale peut être faite par la poste ou par téléphone et être parfois automatisée. Dans les sections 2, 3 et 4, nous analysons le sous-échantillonnage de la population formée des $(n - n_1)$ non-répondants, tel que l'ont proposé Hansen et Hurwitz (1946). Selon cette méthode, la population est censée être constituée d'une strate de réponse de taille N_1 et d'une strate de non-réponse de taille $N_2 = (N - N_1)$.

Dans la section 2, nous analysons deux méthodes permettant de stratifier a posteriori les unités échantillonnées avant de procéder au sous-échantillonnage des non-répondants. Dans la section 3, nous considérons deux estimateurs par quotient de la moyenne d'une caractéristique. Les biais et les erreurs quadratiques moyennes de ces estimateurs sont comparés dans les sections 3 et 4. Dans cette dernière section, nous proposons deux autres estimateurs par quotient qui peuvent convenir dans certaines situations, et analysons leurs avantages relatifs.

La question des cas problèmes est traitée dans la section 5. Nous proposons à cet égard six estimateurs. Nous décrivons aussi brièvement les conditions optimales d'utilisation de chacun de ces estimateurs.

2. ESTIMATEUR DE HANSEN ET HURWITZ ET STRATIFICATION A POSTERIORI

Prenons la caractéristique d'intérêt y_i , $i = (1, 2, \dots, N)$. Définissons par $\bar{Y} = (\sum_{N_1}^1 y_i) / N$ et par $S^2 = \sum_{N_1}^1 (y_i - \bar{Y})^2 / (N - 1)$ la moyenne et la variance de la population. Définissons par $\bar{Y}_1 = (\sum_{N_1}^1 y_i) / N_1$ et par $S_1^2 = \sum_{N_1}^1 (y_i - \bar{Y}_1)^2 / (N_1 - 1)$ la moyenne et la variance du groupe de réponse. De même, nous dirons que $\bar{Y}_2 = (\sum_{N_2}^1 y_i) / N_2$ et $S_2^2 = \sum_{N_2}^1 (y_i - \bar{Y}_2)^2 / (N_2 - 1)$

¹ P.S.R.S. Rao, Département de statistique, Université de Rochester, Rochester, New York 14627, E.-U.

2. $V_2(\hat{t}) = 0$ si le taux de non-réponse est nul ($r = s$);

3. $V_2(\hat{t})$ est réduite sensiblement si nous disposons d'une covariable forte; celle-ci n'a toutefois aucun effet sur $V_1(\hat{t})$ et avec raison puisqu'elle n'est observée que pour $k \in s$.

Regardons de plus près les estimateurs de la variance. Si $V'_i(\hat{t})$ désigne l'estimateur de $V'_i(\hat{t})$, $i = 1, 2$, la variance totale $V(\hat{t})$ peut être estimée au moyen d'une expression du genre

$$V(\hat{t}) = V_1(\hat{t}) + V_2(\hat{t}).$$

Dans l'équation ci-dessus, la variance d'échantillonnage estimée est

$$V_1(\hat{t}) = \sum_{k \in r} \sum_{l \in r} \left(\frac{\pi_k \pi_l}{1} - \frac{\pi_{kl}}{1} \right) \frac{1}{\pi_{kl|s,\bar{m}}} u_k u_l,$$

où $\pi_{kl|s,\bar{m}}$ est définie par (3.2), et π_k, π_{kl} sont les probabilités d'inclusion du plan de sondage. La variance de non-réponse estimée est

$$V_2(\hat{t}) = \sum_{h=1}^H n_h^2 \left(\frac{1}{m_h} - \frac{1}{S_{2wrh}} \right)$$

où

$$S_{2wrh}^2 = \frac{1}{1 - m_h} \sum_{r_h} (w_k - w_{rh})^2.$$

Les quantités u_k et w_k varient d'un estimateur \hat{t} à un autre. Examinons tout d'abord la variance de non-réponse estimée $V_2(\hat{t})$. La définition de cette composante rappelle un échantillonnage stratifié. En effet, le facteur $n_h^2 (1/m_h - 1/n_h)$ est une expression qui caractérise un échantillonnage aléatoire simple stratifié suivant lequel m_h unités sont prélevées parmi n_h unités dans la h -ième strate. Cette structure s'explique par les probabilités de réponse conditionnelles $\pi_{kl|s,\bar{m}}$ définies en (3.2).

Les quantités w_h sont définies de la façon suivante:

$$\text{Pour } \hat{t}_{\text{EXP}} \text{ et } \hat{t}_{\text{EXP}}^*: w_k = \frac{\pi_k}{y_k - \bar{y}_r},$$

$$\text{pour } \hat{t}_{\text{RA}} \text{ et } \hat{t}_{\text{RA}}^*: w_k = \frac{\pi_k}{y_k - (\bar{y}_r / \bar{x}_r) x_k},$$

$$\text{pour } \hat{t}_{\text{REG}} \text{ et } \hat{t}_{\text{REG}}^*: w_k = \frac{\pi_k}{y_k - \bar{y}_r - b(x_k - \bar{x}_r)}.$$

Les expressions ci-dessus sont des valeurs résiduelles de régression affectées d'un poids de sondage. Par conséquent, si x_k est une variable explicative puissante pour y_k , la variance de w_k (et, par voie de conséquence, $V_2(\hat{t})$) sera moins élevée pour les estimateurs de type RA et REG que pour l'estimateur de type EXP, où la quantité w_k n'est que l'écart entre y_k et la moyenne de l'ensemble de réponse \bar{y}_r . Ainsi, dans des conditions favorables, la portion de l'erreur type qui est attribuable à la non-réponse tendra vers zéro, notamment lorsqu'il aura une corrélation quasi parfaite entre x et y .

l'intérieur des groupes. Et même si elle l'était, les groupes sont peut-être mal définis, c'est-à-dire que le modèle aurait du prévoir un plus grand nombre de groupes, etc. Pour la suite de notre analyse, il est donc nécessaire de distinguer deux situations:

- Le mécanisme de réponse hypothétique (MRH, en l'occurrence le modèle A) est exact (en pratique, il est peu probable qu'il en soit réellement ainsi).
- Le MRH est plus ou moins fondé. C'est ce qu'on observe malheureusement dans la plupart des cas et il en découle un biais dû à la non-réponse. En ce qui concerne le modèle A, les groupes peuvent être formés plus ou moins correctement.

Comme cela se fait couramment en statistique, l'analyste définira son modèle en fonction de ce qu'il croit être les meilleurs critères; par conséquent, il fera certaines inférences (jugements de confiance par exemple). Il s'interrogera ensuite sur la robustesse de ses conclusions, c'est-à-dire qu'il cherchera à tester leur validité dans le cas où le modèle ne serait pas fondé. Dans le même ordre d'idées, appliquons ces questions au cas qui nous occupe.

4. ESTIMATEURS DE LA VARIANCE FONDÉS SUR UN MÉCANISME DE RÉPONSE HYPOTHÉTIQUE DONNÉ

Etant donné un ensemble de groupes précis, nous supposons que le modèle A est fondé. Les taux de réponses, $f_h = m_h/n_h$ ($h = 1, \dots, H$), ont été déterminés. À l'aide de ces renseignements préliminaires, nous allons étudier les estimateurs de la variance qui doivent servir à construire un intervalle de confiance à un niveau de $100(1 - \alpha)\%$. Si f est un des estimateurs définis à la section 2 et que le modèle A est vraiment exact, nous constatons ce qui suit:

- f est non biaisé (sauf pour un biais technique habituellement négligeable)
- un intervalle de confiance pour f à un seuil d'environ $100(1 - \alpha)\%$ est

$$f \pm z_{1-\alpha/2} \sqrt{V(f)},$$

où la probabilité que la variable unitaire normale soit supérieure à la constante $z_{1-\alpha/2}$ est de $\alpha/2$. Par des prélèvements successifs d'échantillons s et la formation répétée d'ensembles de réponses r pour chaque échantillon permanent s (conformément au modèle hypothétique A) l'intervalle contiendra le total réel de la population dans $100(1 - \alpha)\%$ des cas. La variance et sa valeur estimée seront déterminées par deux séries de probabilités de sélection:

- π_k et π_{kl} , les probabilités d'inclusion (du premier et du second ordre) qui se rattachent à l'étape de l'échantillonnage;
- $\pi_{k|s,\bar{m}}$, $\pi_{kl|s,\bar{m}}$ les probabilités de réponse conditionnelles (du premier et du second ordre) qui se rattachent au modèle A (étape de la non-réponse).

Dans le cas qui nous occupe et en raison du modèle A, les probabilités $\pi_{k|s,\bar{m}}$, et $\pi_{kl|s,\bar{m}}$, sont définies respectivement par (3.1) et (3.2). En ce qui concerne π_k et π_{kl} , nous supposons qu'elles ont un caractère tout à fait général; nous pouvons utiliser n'importe quel plan de sondage.

Par une analyse détaillée, nous constaterons que la variance totale de n'importe quel estimateur f définit à la section 2 se divise en deux éléments:

$$V(f) = V_1(f) + V_2(f)$$

où $V_1(f)$ désigne la variance d'échantillonnage et $V_2(f)$ la variance due à la non-réponse. Bien que les formules exactes, définies dans Särndal et Swensson (1985a), ne soient pas reproduites ici, on constate que ces deux éléments ont des propriétés valables:

- $V_1(f) = 0$ si l'ensemble de la population U est observée (c'est-à-dire, s'il s'agit d'un recensement plutôt que d'une enquête par sondage);

3. MODÈLES DE RÉPONSE

Les poids de correction pour la non-réponse servant au calcul des estimateurs décrits dans la section précédente peuvent être justifiés au moyen d'un modèle de réponse qui prévoit une probabilité de réponse constante pour chaque unité d'un groupe donné. En termes plus formels, examinons le mécanisme de réponse suivant.

MODÈLE A :

- 1) probabilité de réponse est constante (et égale à une valeur inconnue Θ_h) pour toutes les unités $k \in s_h; h = 1, \dots, H;$
- 2) les unités répondent indépendamment l'une de l'autre.

Les probabilités de réponse théoriques Θ_h peuvent varier considérablement d'un groupe à un autre. (Le fait qu'il peut exister de fortes différences de probabilité de réponse entre divers sous-ensembles porte évidemment à créer des groupes de correction et à faire les pondérations nécessaires.)

Considérons un échantillon permanent s . Les fréquences de groupe $n_1, \dots, n_h, \dots, n_H$ sont alors fixes. Considérons également une valeur constante pour le vecteur des fréquences de réponse des groupes $\tilde{m} = (m_1, \dots, m_h, \dots, m_H)$. Étant donné que s et \tilde{n} sont fixes, il est possible de montrer que le prélevement d'un ensemble de réponses r_h suivant le modèle A équivaut à l'échantillonnage aléatoire simple où l'on choisit m_h unités de n_h . La probabilité de réponse conditionnelle d'une unité k appartenant au h -ième groupe est donc définie :

(3.1)
$$\pi_{k|s,\tilde{m}} = \frac{m_h}{n_h} = f_h, \text{ pour tout } k \in s_h.$$

(Ce raisonnement est à la base du poids f_h^{-1} utilisé dans le calcul des estimateurs.) De même, étant donné s et \tilde{m} , il est possible de montrer que la probabilité que les unités k et l répondent toutes deux à l'enquête suivant le Modèle A est

(3.2)
$$\pi_{k|l,s,\tilde{m}} = \left\{ \begin{array}{ll} f_h & \text{si } k = l \\ \frac{f_h(m_h - 1)}{n_h - 1} & \text{si } k \neq l \in s_h \\ f_h f_{h'} & \text{si } k \in s_h; l \in s_{h'} (h \neq h') \end{array} \right.$$

(par définition, $\pi_{k|s,\tilde{m}}$ est égale à $\pi_{k|s,\tilde{m}}$). Ces valeurs (qui nous rappellent un échantillonage aléatoire stratifié de m_h unités parmi n_h unités dans la h -ième strate) sont importantes pour le calcul des estimations de la variance et des erreurs types (voir ci-dessous). En pratique, l'analyste décide de la méthode de formation des groupes s_h . Cette décision est cruciale puisqu'elle déterminera les poids de correction f_h^{-1} et, par voie de conséquence, la valeur numérique de l'estimation de t , l'estimation de la variance et l'intervalle de confiance. Deux modes de classement différents peuvent produire des estimations ponctuelles et des intervalles de confiance entièrement différents.

L'analyste n'ira pas s'imaginer que les groupes qu'il vient de créer ont tous la même probabilité de réponse. Il est toutefois convaincu (le plus souvent à juste titre) que ces groupes (et, par conséquent, les poids f_h^{-1}) lui permettront d'obtenir des estimations ponctuelles et des intervalles de confiance plus valables. La méthode des groupes de correction est une méthode fiable et solidement établie.

Après une analyse plus minutieuse, on peut découvrir plusieurs irrégularités dans un modèle de réponse comme le modèle A : la probabilité de réponse n'est peut-être pas constante à

(Note: Les poids de sondage ainsi que les poids de correction pour la non-réponse sont utilisés pour déterminer b .)

En résumé, nous avons trois estimateurs

$$\begin{aligned} (2.2a) \quad \hat{t}_{EXP}^* &= N \bar{y}_r, \\ (2.2b) \quad \hat{t}_{RA}^* &= N \bar{x}_s \frac{\bar{y}_r}{\bar{x}_r}, \\ (2.2c) \quad \hat{t}_{REG}^* &= N \{ \bar{y}_r + b (\bar{x}_s - \bar{x}_r) \}. \end{aligned}$$

Tous les trois sont pondérés convenablement avec des poids de sondage et des poids de correction pour la non-réponse. Le principal élément de distinction entre eux est la covariable: t_{EXP}^* n'en renferme aucune alors que t_{RA}^* et t_{REG}^* en contiennent. De plus, t_{RA}^* dénote clairement une relation sous-jacente entre y et la covariable x sous la forme d'une droite passant par l'origine, la pente de cette droite étant estimée par \bar{y}_r/\bar{x}_r . En ce qui concerne l'estimateur t_{REG}^* , la relation sous-jacente est définie par une droite de régression ayant une ordonnée à l'origine non nulle. Nous approfondirons plus loin le rôle de la covariable.

Si l'on connaît la taille de la population N , il est préférable, en règle générale, de substituer N à \hat{N} dans les trois équations précédentes, ce qui donne

$$\begin{aligned} (2.3a) \quad t_{EXP}^* &= N \bar{y}_r, \\ (2.3b) \quad t_{RA}^* &= N \bar{x}_s \frac{\bar{y}_r}{\bar{x}_r}, \\ (2.3c) \quad t_{REG}^* &= N \{ \bar{y}_r + b (\bar{x}_s - \bar{x}_r) \}. \end{aligned}$$

Pour estimer le total de la population, il faut connaître la valeur de N dans chacun des trois estimateurs ci-dessus, ce qui n'est pas toujours le cas. En revanche, pour estimer la moyenne de la population \bar{Y} , on divise chacune de ces équations par N , ce qui donne les équations simplifiées suivantes

$$\begin{aligned} (2.4a) \quad \hat{\bar{t}}_{EXP} &= \bar{y}_r, \\ (2.4b) \quad \hat{\bar{t}}_{RA} &= \bar{x}_s \frac{\bar{y}_r}{\bar{x}_r}, \\ (2.4c) \quad \hat{\bar{t}}_{REG} &= \bar{y}_r + b (\bar{x}_s - \bar{x}_r). \end{aligned}$$

Intuitivement, les trois séries d'estimateurs (2.2), (2.3) et (2.4) sont facilement concevables puisqu'elles reposent sur des règles de pondération élémentaires et les notions fondamentales du quotient ou de la régression. Il est toutefois plus difficile de déterminer l'effet qu'auront ces séries d'estimateurs sur l'estimation de la variance et la construction d'intervalles de confiance valables. Ces questions sont traitées à la section 4. (Contrairement à ce que pourrait laisser croire la présentation plutôt informelle des séries d'estimateurs (2.2) à (2.4), il ne s'agit pas d'équations "ad hoc" mais plutôt d'équations qui découlent d'une méthode d'estimation générale et formalisée (avec une régression à plusieurs variables) pour deux phases de sélection; à cet effet, voir Särndal et Swensson (1985a). Ce qui est encore plus important, c'est que les estimateurs de la variance et les intervalles de confiance découlent directement de cette théorie.)

pourvu que l'on définisse $N = \sum_s 1/\pi_k$, et

$$\hat{y}_r = \frac{\sum_{h=1}^H \sum_{k=1}^K f_{r,h}^{-1} \pi_k}{\sum_{h=1}^H \sum_{k=1}^K f_{r,h}^{-1} \pi_k} \quad (2.1)$$

Le tilde (\sim) sert à indiquer une statistique moyenne qui est pondérée convenablement. On détermine la statistique \tilde{y}_r , qui est la moyenne d'un ensemble de réponses, en affectant la k -ième unité du poids multiplicatif:

poids de sondage \times poids de correction pour la non-réponse $= \pi_k^{-1} f_{r,h}^{-1}$

pour chaque unité k incluse dans le h -ième groupe de correction.

L'estimateur \hat{f}_{EXP} convient bien en situation de non-réponse puisqu'il tient compte du plan de sondage et qu'il tend à corriger l'effet de la non-réponse. Il est toutefois possible d'améliorer \hat{f}_{EXP} , si l'on dispose de plus d'informations. Supposons qu'une covariable simple (et toujours positive) x soit également observée pour $k \in s$. Sur le modèle de l'estimateur par quotient classique, nous pouvons donc définir

$$\hat{f}_{\text{RA}} = \left(\sum_{X_k} \frac{\pi_k}{s} \right) \frac{\sum_{h=1}^H \sum_{k=1}^K f_{r,h}^{-1} \pi_k}{\sum_{Y_k} \frac{\pi_k}{r_h}} = N \hat{x}_s \hat{x}_r$$

où la statistique moyenne \hat{x}_r est définie par l'équation (2.1) en substituant x_k à y_k , et

$$\hat{x}_s = \frac{\sum_{X_k} \frac{\pi_k}{s}}{\sum_{s} \frac{1}{\pi_k}}$$

Comme elle est définie au niveau de l'échantillon visé s , la statistique moyenne \hat{x}_s n'utilise

que des poids de sondage. (Ce genre de moyenne peut être calculée pour la variable x , qui est observée pour tout $k \in s$, mais ne peut évidemment pas l'être pour la variable y qui est observée uniquement pour $k \in r$.)

Dans le cadre de notre analyse, l'équation classique de l'estimateur par régression s'exprime:

$$\hat{f}_{\text{REG}} = N \{ \tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r) \}$$

où

$$b = \frac{\sum_{h=1}^H \sum_{k=1}^K f_{r,h}^{-1} \pi_k (y_k - \tilde{y}_r) (x_k - \tilde{x}_r) / \pi_k}{\sum_{h=1}^H \sum_{k=1}^K f_{r,h}^{-1} \pi_k (x_k - \tilde{x}_r)^2 / \pi_k}$$

Posons $t = \sum U y_k$ comme le total de population à estimer. (Si A est un ensemble d'unités arbitraire, nous écrirons systématiquement $\sum_A y_k$ au lieu de $\sum^{k \in A} y_k$). Selon la méthode des groupes de correction, l'estimateur régulier de t devient alors:

$$(1.1) \quad \hat{t} = \frac{\sum_{h=1}^H f_h^{-1} \sum_{r_h} \pi_{rk}}{\sum_{y_k} \pi_{rk}}$$

Théoriquement, cette méthode est fondée sur l'hypothèse selon laquelle la probabilité de réponse (inconnue) est la même pour chaque unité du même groupe. (Cette hypothèse est une des conditions du modèle A défini à la section 3 plus loin.) La méthode des groupes de correction exige formellement que chaque unité $k \in s$ soit classée. Les variables (qualitatives) servant à cette classification peuvent donc être considérées comme un moyen d'estimer un mécanisme de réponse fondamental.

Un autre genre de variable peut être observée pour chaque $k \in s$; il s'agit des variables qui expliquent y dans le sens habituel de la théorie de la régression. Ces variables seront appelées covariables. Lorsque celles-ci sont incluses dans un estimateur, non seulement elles réduisent sa variance, mais elles le permettent de mieux contrebalancer le biais dû à la non-réponse. (Les covariables ne sont pas des variables auxiliaires dans le sens normal puisqu'elles ne se rapportent pas à la population globale U mais uniquement à l'échantillon visé s .)

Nous faisons donc une nette distinction dans cette étude entre les deux genres de variables observées pour $k \in s$, soit celles qui servent à estimer le mécanisme de réponse et celles qui expliquent la variable d'intérêt y . Dans la description d'un modèle général pour des données en situation de non-réponse, Little (1983) définit plusieurs genres de variables. Afin de situer notre étude en fonction du modèle de Little, disons que ce que celui-ci appelle l'ensemble des variables de réponse complètes est, dans notre étude, divisé en deux sous-ensembles: un sous-ensemble de variables servant à définir le modèle du mécanisme de non-réponse et un sous-ensemble de variables (les covariables) servant à expliquer la variable de réponse incomplète y . La méthode d'inférence que nous utilisons est la méthode de "quasi-randomisation" (Oh et Scheuren 1983), où "quasi" signifie que l'étape de la sélection du mécanisme de non-réponse doit être définie par un modèle tandis que l'étape de l'échantillonnage est régie par les échantillons.

2. ESTIMATEURS SIMPLES DU TOTAL DE POPULATION CORRIGÉS EN FONCTION DE LA NON-RÉPONSE

En développant un peu l'équation courante (1.1), nous obtenons une autre équation, généralement plus efficace, où les poids de sondage π_k^{-1} ont, semble-t-il, une application plus complète:

$$\hat{t}_{EXP} = \left(\sum_s \pi_s \right) \frac{\sum_{h=1}^H \sum_{f_h^{-1}} \frac{1}{\sum_{r_h} \pi_{rk}}}{\sum_{y_k} \frac{1}{\sum_{f_h^{-1}} \sum_{r_h} \pi_{rk}}}$$

Cette équation, qui se ramène à (1.1) lorsque le plan de sondage est auto-pondéré, peut être exprimée comme la formule développée de la moyenne de l'ensemble de réponses:

$$\hat{t}_{EXP} = N \bar{y}_r,$$

Estimation par la méthode de régression en situation de non-réponse

CARL ERIK SÄRNDAL¹

RÉSUMÉ

Dans les cas de non-réponse totale, deux genres de variables se dégagent parfois des unités de l'échantillon initial s ; ce sont les variables servant à estimer le mécanisme de réponse (les probabilités de réponse) et les variables (en l'occurrence, les covariables) qui expliquent la variable d'intérêt dans le sens normal de la théorie de la régression. Dans le présent article, fondé sur l'ouvrage de Särndal et Swensson (1985 a, b), nous analysons les estimateurs corrigés en fonction de la non-réponse en faisant tantôt intervenir des covariables et tantôt non. Cette étude nous amène à conclure qu'un estimateur qui renferme de fortes covariables tend à présenter plusieurs caractéristiques favorables. Par exemple, les estimateurs qui utilisent des covariables sont beaucoup moins exposés au biais dû à la non-réponse. Nous examinons aussi le calcul d'erreurs types et d'intervalles de confiance valables pour les estimateurs comprenant des covariables. Enfin, nous analysons la structure de l'erreur type.

MOTS CLÉS: Mécanisme de réponse; méthode des groupes de correction; covariable; robuste.

1. INTRODUCTION

Nous considérons une population finie $U = \{1, \dots, k, \dots, N\}$, de laquelle un échantillon s de taille n est prélevé à l'aide d'un plan de sondage selon lequel la probabilité de sélection (strictement positive) de la k -ième unité est π_k . Le poids d'échantillonnage de la k -ième unité est donc π_k^{-1} . Nous pouvons supposer un plan de sondage complexe, qui n'est pas nécessairement auto-pondéré, par exemple, un plan de sondage à 3 degrés avec sélection stratifiée des unités primaires. Selon ce plan, la probabilité que les unités k et l soient toutes deux incluses dans l'échantillon est désignée π_{kl} ($\pi_{kl} > 0$ pour tout $k \neq l$ et π_{kk} est considérée égale à π_k).

Étant donné l'échantillon s , nous supposons qu'il y aura un certain taux de non-réponse totale. Nous désignons r le sous-ensemble de s que constituent les répondants et m la taille de ce sous-ensemble. La variable d'intérêt, y , est observée uniquement pour $k \in r$. Afin de compenser le biais dû à la non-réponse, nous supposons, pour les besoins de la présente étude, que la méthode très connue des groupes de correction est utilisée: l'échantillon s est subdivisé en H groupes $s_1, \dots, s_h, \dots, s_H$ de tailles respectives $n_1, \dots, n_h, \dots, n_H$. De même, nous subdivisons l'ensemble de réponses r en sous-ensembles $r_1, \dots, r_h, \dots, r_H$ de tailles respectives $m_1, \dots, m_h, \dots, m_H$. Le taux de réponse pour le groupe h est défini comme $f_h = m_h/n_h$. La méthode précitée comporte une opération qui consiste à affecter les observations du groupe h d'un poids de correction f_h^{-1} , en plus du poids de sondage. (Dans le présent modèle, nous supposons que la taille et la composition des groupes de correction au niveau de la population sont inconnues.) Nous avons:

$$n = \sum_{h=1}^H n_h; m = \sum_{h=1}^H m_h.$$

¹ Carl Erik Särndal, département de mathématique et de statistique, Université de Montréal, Montréal, Québec), Canada, H3C 3J7.

(iv)
$$Y_4 = \sum_{j=1}^k \frac{N'_k}{m_{1k}} m_{jk}$$

$$V(Y_4) = \sum_{k=1}^K N'_k \left(\frac{m_{1k}}{s_{yk}^2} - m_{1k} \right) s_{yk}^2$$

Des estimateurs par le quotient peuvent être calculés. Comme pour les estimateurs simples, ils peuvent prendre différentes formes, dépendant des hypothèses formulées. Par exemple, l'estimateur par le quotient équivalait à l'estimateur 4 serait:

$$Y_{\hat{Q}4} = \sum_{k=1}^K N'_k Y_{\text{sub}k} \frac{X_{\text{samp}k}}{X_{\text{sub}k}}$$

où $X_{\text{samp}k}$ est la moyenne de la variable X pour les unités sélectionnées dans l'échantillon complet, qui sont dans la strate k
 $X_{\text{sub}k}$ est la moyenne de X pour les unités sélectionnées dans le sous-échantillon, qui sont dans la strate k
 $Y_{\text{sub}k}$ est la moyenne de la variable Y dans la strate k du sous-échantillon.

$$V(Y_{\hat{Q}4}) = \sum_{k=1}^K (N'_k)^2 \left(\frac{1}{m_{1k}} - \frac{1}{s_{yk}^2} \right) \left[s_{yk}^2 + R_{k^2 s_{yk}^2} - 2R_{k s_{y x k}} + \left(\frac{1}{m_{1k}} - \frac{1}{s_{yk}^2} \right) s_{yk}^2 \right]$$

où $R_k = \frac{Y_{\text{sub}k}}{X_{\text{sub}k}}$.

REMERCIEMENTS

L'auteur tient à remercier P. Gile pour les commentaires et les suggestions fournis lors de cette étude, de même que les arbitres pour leurs remarques pertinentes.

BIBLIOGRAPHIE

BANKIER, M., (1982). Variance formula for an estimator based on any number of independent stratified samples of which some are Poisson samples. Document technique, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.

COLLEDGE, M.L., JOHNSTON, J.H., PARÉ, R., et SANDE, I.G. (1978). Large scale imputation of survey data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 721-726.

GILES, P. (1983). Construction division: Census of construction. Document Technique, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

PHILIPS, J.L., et EMERY, D. (1976), FIBCOC documentation Document Technique, Division des Developpements de système, Statistique Canada.

RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

N_{2h} : la taille de la population "pré-spécifiée" avec des entreprises "hors du champs de l'enquête", dans la strate h (estimé),
 N_k : la taille de la population dans la strate k , estimée à partir de l'information de l'échantillon transversal,
 N'_k : la taille de la population dans la strate k , estimée à partir de l'information des deux échantillons (bases multiples),
 n_h : le nombre de données échantillonnées dans la strate h de l'échantillon pré-spécifié,
 n_{1h} : le nombre de données échantillonnées et transcrites de la strate h de l'échantillon pré-spécifié,
 n'_k : le nombre de données échantillonnées et transcrites dans la strate k ,
 m_{1h} : le nombre de données sous-échantillonnées parmi les "vivants" de la strate h ,
 y : une variable d'un des sous-échantillons,
 x : une variable auxiliaire disponible pour toutes les données des échantillons,
 S_{2y}^2 : une estimée de la variance de y pour les données du sous-échantillon dans la strate h ,
 S_{2x}^2 : une estimée de la variance de x pour les données du sous-échantillon dans la strate h ,
 S_{yxh} : une estimée de la covariance entre x et y dans la strate h .

$$I) \quad Y_1 = \left(\frac{N_{1 \text{ pre-spec.}} + N_{\text{births}}}{N_{1 \text{ pre-spec.}}} \sum_{h=1}^h \frac{N_h}{n_{1h}} \frac{m_{1h}}{n_{1h}} \sum_{hj} Y_{hj} \right) \quad (1)$$

$$V(Y_1) \approx \left(\frac{N_{1 \text{ pre-spec.}} + N_{\text{births}}}{N_{1 \text{ pre-spec.}}} \right)^2 \sum_{h=1}^h N_h n_h \left(\frac{n_h - 1}{n_h - 1} \right) \quad (2)$$

$$\times \left[W_{1h} S_{2y}^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) + \frac{n_h}{G_h} S_{2y}^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) + \frac{n_h}{G_h} W_{1h} (1 - W_{1h})^2 S_{2y}^2 \right]$$

$$\text{où } G_h = \left(\frac{N_h - n_h}{N_h - 1} \right), \gamma_h = n_h \frac{n_{1h}}{m_{1h}}, \text{ et } W_{1h} = \frac{n_{1h}}{n_h}.$$

$$II) \quad Y_2 = \left(\frac{N_{1 \text{ pre-spec.}} + N_{\text{births}}}{N_{1 \text{ pre-spec.}}} \sum_{h=1}^h \frac{N_h}{m_{1h}} \sum_{hj} Y_{hj} \right) \quad (2)$$

$$V(Y_2) \approx \left(\frac{N_{1 \text{ pre-spec.}} + N_{\text{births}}}{N_{1 \text{ pre-spec.}}} \right)^2 \sum_{h=1}^h \frac{N_h}{m_{1h}} \sum_{hj} Y_{hj} \quad (2)$$

$$III) \quad Y_3 = \sum_k \frac{N_k}{N_{kj}} \sum_{kj} Y_{kj} \quad (3)$$

$$V(Y_3) = \sum_k N_k \left(\frac{m_{1k}}{N_k - m_{1k}} \right) S_{yk}^2 \quad (3)$$

Tableau 4

Estimés de ADD obtenus par simulation

	Population	Pondération	Ratio	Imputation
estimé ($\times 10^7$)	1.41	1.43	1.24	1.41
écart-type ($\times 10^5$)		1.11	.85	1.15
biais ($\times 10^5$)		.22	-1.73	-0.07

Tableau 5

Estimés de RM obtenus par simulation

	Population	Pondération	Ratio	Imputation
estimé ($\times 10^7$)	1.06	1.06	1.07	1.04
écart-type ($\times 10^5$)		4.52	4.11	4.87
biais ($\times 10^5$)		-0.07	-0.95	-1.38

Pour la variable ADD, l'estimé obtenu par le ratio est significativement différent des estimés obtenus par imputation ou par pondération. Le biais de L'estimé est aussi significativement non nul. Pour la variable RM, tous les estimateurs sont équivalents (variance égales, biais non significatif à un seuil de 5%, estimés non significativement différents).

7. CONCLUSIONS

D'après l'étude, il ne semble pas y avoir de différences significatives entre une stratégie d'imputation massive et de pondération, pour les variables du sous-échantillon autres caractéristiques. Le résultat est prévisible dans la mesure où les variables étudiées semblent relativement stables à l'intérieur des strates.

Les conclusions pour les variables du sous-échantillon financier se basent sur les résultats obtenus par la simulation. De cette étude, il semble que la pondération par l'inverse de la probabilité de sélection et l'imputation massive donnent des estimés comparables. L'estimateur par le ratio ne semble pas approprié pour la variable ADD (ainsi que pour d'autres variables étudiées, non présentées ici). Une autre étude cherchera à étudier si l'estimateur par la régression serait plus approprié. L'étude visera aussi à évaluer l'impact de l'imputation sur la structure de corrélation des variables.

ANNEXE

Les estimateurs proposés pourraient être écrits de la façon suivante:

Notations

- Soit h : les strates de l'échantillon pré-spécifié,
- k : les strates de l'échantillon transversal,
- N_h : la taille de la population "pré-spécifiée" dans la strate h ,
- N_{1h} : la taille de la population "pré-spécifiée" avec des entreprises "vivantes" (dans le champs de l'enquête) dans la strate h (estimé),

Tableau 3

Différentes estimations de RM et de l'écart-type de RM

	Y_1	Y_2	Y_3	Y_4	Y_{Q2}	Y_{Q3}	Y_{Q4}	Y_1
estimation ($\times 10^8$)	1.5	1.5	1.43	1.55	0.9	1.63	1.67	1.75
écart-type ($\times 10^6$)	6.9	6.9	8.9	5.3	3.1	11.0	4.3	

On peut remarquer dans un premier temps, qu'il n'y a pas de différences significatives entre les deux premiers estimateurs. (Selon les définitions préalablement établies le second estimateur est une version simplifiée du premier estimateur). La version simplifiée sera donc conservée.

Pour les variables dans le sous-échantillon non financier, en général, l'imputation semble donner des résultats semblables à ceux obtenus par la pondération (Y_4). L'estimateur obtenu en ne considérant que les unités provenant de l'échantillon transversal (Y_3) semble plus variable que les autres estimateurs. Cette variabilité pourrait être expliquée par le plus petit nombre d'unités utilisées dans le calcul de cet estimateur. Il est à noter que ces observations ne sont basées que sur un échantillon observé et que les conclusions sont donc limitées. Cependant, à cause de la nature des données (souvent des pourcentages et des subdivisions d'activité dans le domaine de la construction), qui est relativement stable dans les strates (CAB 1970 3 chiffres, province et RB) il n'a pas été jugé nécessaire de pousser plus à fond l'étude de ces variables.

Pour les variables dans le sous-échantillon financier il a été remarqué qu'en général, les estimateurs ajustés par le quotient ne semblent pas toujours applicables (ex: variable ADD). Ils donnent des estimés extrêmement biaisés. Une explication possible est que la variable ADD et la variable auxiliaire utilisée ont une grande fréquence de zéro. Un "mauvais" échantillon dans certaines strates peut donc faire gonfler excessivement les estimés. Certains problèmes ont aussi été observés avec le système d'imputation, (données imputées alors qu'elles n'auraient pas dû, données non imputées), ce qui a parfois pu influencer les estimés obtenus par la stratégie d'imputation. Parce que les résultats étaient basés sur un échantillon observé seulement et parce qu'il était difficile d'estimer l'impact des problèmes reliés au système, sur les estimés, il a été décidé de faire une simulation.

6. SIMULATION

La simulation a été effectuée sur un sous-ensemble de données; soit les entreprises ayant été sélectionnées dans le sous-échantillon financier (pour ce sous-ensemble de données, toutes les variables étudiées sont présentes). Par la suite, il a été essayé de reprendre de façon simplifiée la stratégie utilisée par le recensement de la construction. Un échantillon stratifié a été sélectionné. Des fractions de sondage semblables à celles de l'enquête ont été utilisées dans l'échantillonage. Les variables du sous-échantillon financier, pour les données non-sélectionnées dans l'échantillon, ont été mises manquantes, pour être ensuite imputées par le système. Le processus de sélection d'un échantillon, puis d'imputation a été répété trente fois. Des estimés ont été produits, comparant les résultats obtenus en sommant les données non-imputées et imputées, par rapport aux estimés obtenus en pondérant l'échantillon par l'inverse de la fraction de sondage. Comme la valeur de la population est connue, le biais a été calculé, en plus de la variance des estimés. Les résultats pour les variables ADD et RM sont présentés dans les tables 4 et 5.

Comme mentionné précédemment, les variables recueillies dans le sous-échantillon financier sont ajustées par le quotient d'une variable auxiliaire lors de l'imputation. Pour cette raison, un autre type d'estimateur pourrait être envisagé pour les variables recueillies dans le sous-échantillon financier: soit un estimateur par le quotient. La variable auxiliaire utilisée serait la variable utilisée lors de l'imputation. Comme pour la pondération simple, différents estimateurs pourraient être calculés.

Les différents estimateurs et leur variance sont écrits de façon mathématiques dans l'annexe 1.

5. RÉSULTATS

Pour l'étude certaines variables ont été choisies parmi chacun des sous-échantillons. Le sous-échantillon financier est composé de sept variables. L'étude s'est limitée à quatre de ces sept variables.

Pour le sous-échantillon autres caractéristiques, huit variables sont recueillies pour toutes les entreprises. Les autres variables sont disponibles pour différents groupes d'activité économique seulement. L'étude se limitera donc à ces huit variables.

Pour le sous-échantillon financier, les variables présentées dans ce rapport sont ADD (additions immobilisations) et RM (réparation et entretien). Pour le sous-échantillon A.C., la variable PCON (pourcentage de construction dans un domaine particulier) est présentée. Cependant, la variable PCON n'est pas publiée directement. Elle est multipliée par les dépenses totales, pour obtenir les dépenses dans un domaine particulier: PEXP. C'est cette deuxième variable qui a été étudiée.

Comme mentionné précédemment, les variables dans le sous-échantillon A.C. ne sont pas ajustées par un quotient, lors de l'imputation. Les estimateurs par le ratio ne s'appliqueront donc pas à ces variables.

Les tables 1, 2 et 3 nous présentent les différents estimateurs et les estimateurs de leurs variances respectives (basé sur les données fiscales de 1983, pour les entreprises non incorporées).

Tableau 1
Différentes estimations de PEXP (%EXP*EXPCONS)

estimation (× 10 ¹¹)	3.44	3.43	3.5	8.4	3.66	3.2
écart-type (× 10 ⁹)	3.5	3.5	3.5	8.4	3.66	3.2
Y ₁	Y ₂	Y ₃	Y ₄	Y ₁	Y ₂	Y ₃

Tableau 2
Différentes estimations de ADD et de l'écart-type de ADD

estimation (× 10 ⁸)	2.08	2.10	2.14	1.84	7.82	5.06	5.2	1.4
écart-type (× 10 ⁷)	1.9	1.9	2.0	1.0	0.8	2.2	0.8	0.8
Y ₁	Y ₂	Y ₃	Y ₄	Y _{Q2}	Y _{Q3}	Y _{Q4}	Y ₁	Y ₁

Les naissances et les morts ne peuvent être classifiés de façon croisée. Les morts ont un poids $W_h = 0$ et les naissances, un poids W_k inversement égal à la probabilité de sélection dans la strate k de l'échantillon transversal. Plus de détails peuvent être obtenus dans Bankier (1982).

L'estimateur du total obtenu, lorsque la stratégie d'imputation est utilisée, est donc

$$Y = \sum_{h,k} W_{hk} \sum_{j=1}^n Y_{jk}^*$$

où $Y_{jk}^* = Y_{jk}$ si $j \in$ sous-échantillon

$$= Y_{jhk}^I \text{ si } j \notin \text{ sous-échantillon.}$$

4. STRATÉGIE DE PONDERATION

Si une stratégie de pondération était utilisée pour estimer les variables des sous-échantillons, différents estimateurs pourraient être envisagés. Les estimateurs ont la même forme pour les deux sous-échantillons. Cependant les poids utilisés sont différents.

Un premier estimateur (Y_1) serait un estimateur basé sur le plan d'échantillonnage utilisé, ajusté pour la sous-couverture de la population. Dans chaque strate de CAE, PROV et RB (code d'activité économique, provenance, revenu brut), un échantillon pré-spécifié est sélectionné. Lorsque transcriptes (unités échantillonnées et encore vivantes), les unités sont classées en deux strates : "hors du champs de l'enquête" et "dans le champs de l'enquête". Les sous-échantillons sont choisis à partir de la strate d'unités "hors du champs de l'enquête". (On peut supposer que toutes les unités de la strate "hors du champs de l'enquête" ont été sous-échantillonnées et qu'elles ont une moyenne égale à zéro). L'estimateur contient un facteur de correction, pour compenser pour la sous-couverture de la base de sondage (obtenu à partir de l'information de l'échantillon transversal).

Un second estimateur possible (Y_2) serait une version simplifiée du premier estimateur proposé. Au lieu de supposer un double échantillonnage pour déterminer les unités "dans le champs de l'enquête" et "hors du champs de l'enquête", on pourrait supposer qu'un échantillon stratifié pré-spécifié est choisi à partir des unités "dans le champs de l'enquête". Un sous-échantillon est sélectionné à partir de l'échantillon pré-spécifié. L'estimateur doit encore une fois, être ajusté pour tenir compte de la sous-couverture. S'il ne s'avère pas y avoir de différences significatives entre le premier et le second estimateur, le second serait préférable, parce que plus simple.

Un troisième estimateur possible (Y_3) serait un estimateur basé sur l'information provenant de l'échantillon transversal et dans l'échantillon et le sous-échantillon. On pourrait supposer que les unités choisies à la fois dans le sous-échantillon et dans l'échantillon transversal ont été choisies à partir de l'échantillon transversal. Le rationnel pour un tel estimateur serait que l'échantillon transversal est sélectionné à partir d'une base de sondage complète. Cependant, comme les sous-échantillons sont sélectionnés à partir de l'échantillon pré-spécifié et non à partir de l'échantillon transversal, la taille des sous-échantillons dans l'échantillon transversal sera petite. Finalement un dernier estimateur possible (Y_4) serait de supposer que le sous-échantillon a été sélectionné à partir de l'échantillon complet (échantillon pré-spécifié + échantillon transversal), et que l'échantillon complet provient de bases multiples. Ce quatrième estimateur est celui qui est le plus "semblable" à l'estimateur obtenu après l'imputation massive. En effet, ces deux estimateurs supposent que les naissances et les nouvelles entreprises "réagissent" comme le reste de la population. Aucun ajustement particulier n'est fait pour ces entreprises, dans la technique d'imputation, et l'estimateur pondéré n'est pas stratifié de façon à distinguer ces unités. De plus les deux estimateurs tiennent compte du fait que l'échantillon provient de bases multiples. Le même poids est donc utilisé dans les deux cas, pour pondérer l'échantillon à la population.

3. STRATÉGIE D'IMPUTATION

Le RC utilise une stratégie d'imputation massive, pour estimer les variables sélectionnées dans un sous-échantillon donné (i.e. que pour tous les enregistrements non sélectionnés dans les sous-échantillons, des valeurs sont imputées pour chaque variable). Le processus d'imputation se fait de façon indépendante pour chaque sous-échantillon. (La procédure d'imputation se fait par phases. Les phases d'imputation des sous-échantillons sont indépendantes entre elles et elles utilisent des stratégies différentes). Dans chaque phase, le plus proche voisin est sélectionné comme donneur pour imputer les variables non-échantillonnées (le plus proche voisin est choisi parmi un sous-groupe de donneurs potentiels).

L'imputation se fait de façon différente pour chaque sous-échantillon.

Pour le sous-échantillon financier, la valeur imputée est la valeur du donneur, ajustée par le quotient d'une variable auxiliaire, disponible à la fois pour le donneur et pour le candidat (le candidat étant défini comme étant l'unité ayant besoin d'être imputée). (Note: La procédure réelle est plus complexe qu'une simple imputation ajustée par le quotient. En effet, les variables sont imputées de façon hiérarchique et des contraintes linéaires sont imposées aux valeurs imputées (la deuxième variable réelle peuvent être trouvés dans Phillips et Emery (1976). Une vue générale plus détaillée se trouve aussi dans Colledge et coll (1978)).

Soit Y: la variable d'intérêt (qui doit être imputée pour le candidat, connue pour les donneurs),

X: une variable auxiliaire disponible à la fois pour le donneur et le candidat,

c: un indice référant au candidat,

d: un indice référant au donneur,

I: un indice référant à une valeur imputée.

Pour les variables du sous-échantillon financier la valeur imputée Y^c_c est définie comme étant:

$$Y^c_c = Y^d_c \frac{X^d_c}{X^c_c}$$

Pour les variables du sous-échantillon A.C., la valeur imputée est la valeur du donneur.

$$Y^c_c = Y^d_c$$

Après l'étape d'imputation, il existe un fichier rectangulaire complet (toutes les entreprises ayant été sélectionnées dans un des échantillons possèdent des valeurs pour toutes les variables des échantillons sous-échantillons). En pondérant l'échantillon, des estimés au niveau de la population peuvent être obtenus.

Le poids utilisé peut être décrit comme étant l'inverse de la probabilité de sélection, dans au moins un des échantillons.

Soit: $P(\text{presp}_h)$: la probabilité qu'une unité ait été sélectionnée dans la strate h de l'échantillon pré-spécifié.

$P(\text{trans}_k)$: la probabilité qu'une unité ait été sélectionnée dans la strate k de l'échantillon transversal.

hk : classification croisée des unités.

h : l'indice relié à la stratification de l'échantillon pré-spécifié.

k : l'indice relié à la stratification de l'échantillon transversal.

Le poids associé à chaque unité peut être écrit comme

$$W^{-1}_{hk} = 1 - [1 - P(\text{presp}_h)] [1 - P(\text{trans}_k)]$$

incorporées sera modifiée pour l'année fiscale 1984 et deviendra équivalente à celle des entreprises non incorporées. C'est donc la stratégie des entreprises non incorporées qui a été étudiée. Il est espéré que les conclusions resteront semblables, pour les entreprises incorporées.

2. DESCRIPTION DU PLAN DE SONDAGE

Comme mentionné précédemment deux échantillons sont sélectionnés de façon indépendante, à partir de deux bases de sondage qui se chevauchent. Un premier échantillon, l'échantillon pré-spécifié, est un échantillon stratifié (par revenu brut (RB), province et code d'activité économique CAE 3 chiffres 1970), sélectionné pour le recensement de la construction. Il est sélectionné à partir d'une base de sondage qui n'est pas complètement à jour. En effet, la base de sondage contient des entreprises "mortes", i.e. des entreprises qui ne sont plus dans le champs de l'enquête pour construction (l'entreprise n'existe plus, elle n'est plus dans un domaine d'activité de construction ou l'entreprise a un revenu brut inférieur à \$10,000). De la même façon, la base de sondage ne contient pas les naissances et les entreprises qui ont changé de domaine d'activité et qui sont maintenant en construction. De façon indépendante, un échantillon "transversal" est choisi par Revenu Canada. Cet échantillon stratifié (par des classes de Revenu Brut) est choisi parmi tous les domaines d'activité économique (pas seulement construction), à partir d'une base de données complète il sert à couvrir les naissances. La figure 1 illustre la situation.

À partir des unités de l'échantillon pré-spécifié, deux sous-échantillons sont sélectionnés de façon indépendante: un sous-échantillon financier et un sous-échantillon "autres caractéristiques" (A.C.). Le sous-échantillon A.C. est sélectionné directement à partir de l'échantillon pré-spécifié, alors que le sous-échantillon financier est sélectionné à partir des données transcrites de l'échantillon (on ne sous-échantillonne donc pas les "morts"). Plus de détails sur le plan de sondage peuvent être obtenus dans Giles(1983).

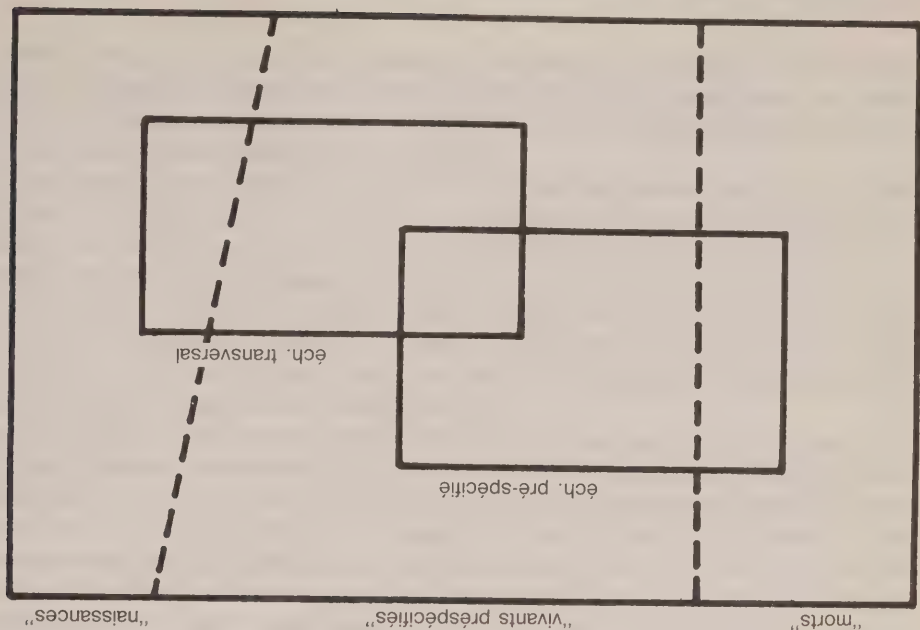


Figure 1. Schéma du plan de sondage du RC

Comparaison de la pondération et de l'imputation pour des données non-échantillonnées

SYLVIE MICHAUD¹

RÉSUMÉ

Le Recensement de la Construction (RC) au Canada se sert d'un plan de sondage complexe pour échantillonner les petites entreprises (les entreprises dont le revenu brut est inférieur à \$750,000). Des échantillons stratifiés sont sélectionnés à partir de base de sondage qui se chevauchent. À partir d'un des échantillons, deux sous-échantillons sont sélectionnés de façon indépendante. De l'information plus détaillée est recueillie, pour les entreprises choisies dans les sous-échantillons. Deux stratégies pourraient être envisagées, pour estimer des totaux pour les variables recueillies dans les sous-échantillons. La première approche serait de déterminer des poids, basés sur les fractions de sondage. Cette approche nécessite l'utilisation de plusieurs poids différents. Une seconde approche serait d'imputer des valeurs aux entreprises sélectionnées dans l'échantillon mais pas dans les sous-échantillons. Cette approche crée un fichier "rectangulaire" complet au niveau de l'échantillon. Un seul poids peut ensuite être utilisé pour obtenir des estimés pour la population. L'étude vise à comparer les estimés qui pourraient être obtenus, par le Recensement de la Construction. L'étude vise à comparer les estimés qui pourraient être obtenus, en utilisant diverses stratégies d'estimation aux estimés obtenus lorsque l'approche d'imputation massive est employée.

MOTS CLÉS: Pondération; imputation massive; non-échantillonné.

1. INTRODUCTION

Le recensement de la construction (RC) est une enquête annuelle qui vise à estimer les dépenses dans le domaine de la construction. Bien que l'enquête porte le titre de "recensement", seules les entreprises ayant un revenu brut supérieure à \$750,000 sont recensées. On envoie un questionnaire long à ces entreprises, dans lequel on recueille diverses informations financières et non financières. On estime les dépenses des entreprises dont le revenu brut se situe entre \$10,000 et \$750,000 à partir d'un échantillon de données administratives. Dans un premier temps, deux échantillons sont sélectionnés de façon indépendante, à partir de bases de sondage qui se chevauchent. Deux sous-échantillons sont ensuite choisis parmi un des échantillons, pour obtenir de l'information supplémentaire.

Deux stratégies pourraient être employées pour estimer des variables recueillies dans les sous-échantillons. Présentement, l'approche utilisée par le recensement de la construction est d'imputer des valeurs pour les entreprises non sélectionnées dans un sous-échantillon, mais choisies dans un échantillon. Cela crée un fichier "rectangulaire" complet, qui peut être pondéré au niveau de la population en n'utilisant qu'un seul poids. Une alternative serait de calculer des poids, basés sur les probabilités de sélection. Différents poids devraient être calculés pour différents sous-ensembles de données. Le but de cette étude est de comparer les estimés obtenus par pondération aux estimés obtenus par imputation.

L'étude est effectuée sur une population d'entreprises non incorporées seulement parce que pour l'année fiscale 1983, les stratégies de sélection des échantillons d'entreprises non incorporées et incorporées étaient différentes. Cependant, la stratégie utilisée pour les entreprises

¹ S. Michaud, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11^{ème} étage, R.H. Coats, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

- Hinkins et Scheuren: Hot deck et l'échantillonnage à deux degrés
- HINKINS, S. (1984). Matrix sampling and the related imputation of corporate income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 415-420.
- JONES, H., et McMAHON, P. (1984). Sampling corporation income tax returns for statistics of income, 1951 to present. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 437-442.
- LESZCZ, M.R., OH, H.L., et SCHEUREN, F.J. (1983). Modified raking estimation in the Corporate SOI Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 434-438.
- LITTLE, R.J.A. (1986). Missing data in Census Bureau surveys. Présenté en mars 1986 lors de la deuxième conférence annuelle relative aux recherches sur le recensement (Census Research Conference). Cet article va paraître dans le *Journal of Business and Economic Statistics*.
- OH, H.L., et SCHEUREN, F.J. (1980). Estimating the variance impact of missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 416-420.
- POWELL, W.T., et STUBBS, J.R. (1981). Using business master file data for statistics of income purposes. *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. 1., Washington, DC: Internal Revenue Service, 157-167. Voir tout particulièrement l'annexe préparée par Alan Freiden.
- RUBIN, D., et SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- SANDE, I.G., (1982). Imputation in surveys: coping with reality. *The American Statistician*, 36, 145-152.
- STRUDLER, M., OH, H.L., et SCHEUREN, F.J. (1986). Protection of taxpayer confidentiality with respect to the IRS individual tax model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (en cours de publication).

pour chaque case appropriée. Si on se reporte au tableau A, on pourrait calculer approximativement l'augmentation de variance V_2 en utilisant l'estimation par la méthode du quotient et les formules de cette méthode (voir Cochran 1977). Toutefois, ces formules sont des approximations faites à l'égard de très grands échantillons tandis que nos échantillons sont presque toujours très petits. Dans le présent cas, la dimension de l'échantillon correspond au nombre de donneurs, n_B , dans une case d'ajustement. Par conséquent, il nous faut ici des résultats empiriques.

De même, il est possible de trouver la distorsion b_2 , en utilisant les résultats de l'estimation par la méthode du quotient. Contrairement aux résultats obtenus par la méthode hot deck, la distorsion obtenue par la méthode du quotient se rapproche de zéro à mesure que la dimension de l'échantillon augmente et, dans ce sens, l'estimation par la méthode du quotient est plus certaine. En fait, l'estimation par la méthode hot deck est non biaisée uniquement si le modèle $Y = \beta X$ est exact. Il est bien entendu que la distorsion se rapproche de zéro dans les deux méthodes d'estimation à mesure que la fraction de données manquantes se rapproche aussi de zéro. Toutefois, même si le modèle $Y = \beta X$ est inexact, l'estimation par la méthode du quotient demeure constante.

Il existe évidemment de nombreuses autres possibilités; on pourrait étudier les modèles de régression à plusieurs variables. Nous sommes encore aux premières étapes de ce projet et nous avons du pain sur la planche pour l'immédiate et pour les années qui viennent.

BIBLIOGRAPHIE

- AMBROSE, P. (1985). Tax year 1985 business finance (T2) sample selection: detailed statement of requirements. Statistique Canada (non publié).
- BARKER, D., HINKINS, S., et REHULA, V. (1982). 1981 corporation validation tests. Statistics of Income Division, Internal Revenue Service (non publié).
- BURPEE, J., et MCGRATH, A. (1982). Micro-model of corporation taxation sample design and estimates. Services statistiques, Revenu Canada-Impôt, (non publié).
- CLICKNER, R.P., GALFOND, G.J., et THIBODEAU, L.A. (1984). Evaluation of the IRS corporate SOI sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 443-448.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e éd.). New York: John Wiley and Sons, Inc.
- COLLEDGE, M., JOHNSON, J., PARE, R., et SANDE, I.G. (1978). Large scale imputation of survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 431-436. (Voir aussi l'article de S. Michaud dans ce numéro.)
- CYS, K., HINKINS, S., et REHULA, V. (1982). Automatic and manual edits for corporation income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 443-448.
- CZAJKA, J. (1986). Imputation of selected items in corporate tax data: improving upon the earlier hot deck. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (en cours de publication).
- FORD, B.L. (1983). An overview of hot deck procedures. Dans *Incomplete Data in Sample Surveys*, Volume 2 - Theory and Bibliographies (éd. W.G. Madow, I. Oikin, et D.B. Rubin), New York: Academic Press, 185-207.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory*, Vol. II, New York: John Wiley and Sons, Inc.
- HINKINS, S. (1983). Matrix sampling and the related imputation of corporate income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 427-433.

Par conséquent, le prix payé pour l'économie réalisée lorsqu'on n'effectue pas le contrôle de chaque annexe est l'augmentation de la variance due au double échantillonnage. Cette augmentation de variance semble potentiellement dommageable parce que K est très grand. Toutefois, il faut se rappeler que $Z = X - Y$ et que l'augmentation de la variance est fonction uniquement de la variance de Y pour la sous-population B . Nous prévoyons que $S^2(X)$ va dominer $S^2(Y)$ qui à son tour devrait dominer $S^2_B(Y)$. En d'autres mots,

$$S^2(X) >> S^2(Y) >> S^2_B(Y),$$

parce que la dimension de la variance est fonction de la valeur moyenne et que Y devrait être petit comparativement à X . Pour la plupart des éléments, nous prévoyons que le montant des erreurs de classification sera petit comparativement au montant original. Par conséquent, nous croyons que $S^2_B(Y)$ sera beaucoup plus petit que $S^2(Z)$ et que $P_B(K - 1)S^2_B(Y)$ demeurera relativement petit comparativement à $S^2(Z)$; l'augmentation de la variance due au sous-échantillonnage sera donc relativement petite. Ce facteur n'est pas garanti, mais les résultats obtenus par Czajka corroborent ce fait pour la grande majorité des éléments (Czajka 1986). *Imputation par la méthode "hot deck"* – On a fait appel à l'imputation par la méthode "hot deck" à l'intérieur des cas de l'ajustement pour reconstruire un ensemble de données rectangulaire. Plus précisément, on a associé une déclaration dont les annexes devaient être imputées à un donneur du groupe B dont les mêmes annexes avaient été contrôlées dans la même case d'ajustement.

En imputant les valeurs manquantes de y par la procédure hot deck et en utilisant un échantillon aléatoire simple, on augmente encore la variance comparativement à celle obtenue lorsqu'on utilise l'estimateur de double échantillonnage (\hat{z}_d). Toutefois, l'augmentation ajoutée à l'augmentation due à l'imputation par la méthode hot deck est petite comparativement à l'augmentation due au double échantillonnage. Cette relative augmentation de la variance due à l'imputation, désignée par c_2 dans le tableau A, est limitée et, dans notre cas, petite. (Lorsque $K \geq 2$, $c_2 \leq 0.125$; à ce sujet, voir par exemple Hansen, Hurwitz et Madow 1953).

Comme nous l'expliquons dans le document, l'emploi de la méthode hot deck ordinaire crée un problème. Si on évalue simplement la valeur y_i non observée à l'égard de l'article i en partant de la valeur observée y_j provenant de l'article donneur j , l'estimation de la valeur z_i ainsi obtenue peut ne pas réussir les contrôles de validation. Il faudrait alors apporter des corrections supplémentaires à l'article. Puisque le montant original est toujours observé, il semblait plus raisonnable de traiter par la méthode "hot deck" le changement relatif $R = Y/X$ au lieu du changement avec rapport devant réduire la variance des estimations comparativement à la variance obtenue par l'approche hot deck de base; toutefois, la variance d'une estimation ne peut être établie par analyse; elle doit être mesurée empiriquement. En outre, lorsqu'on introduit un rapport, les estimateurs deviennent biaisés. Nous avons supposé que les biais ainsi créés seraient petits, comme ils l'ont été dans la plupart des cas ainsi qu'on peut le constater en analysant le tableau 2. En pratique, l'imputation hot deck a été effectuée à l'intérieur des cas d'ajustement créés par post-stratification des articles en cas homogènes, du moins nous l'espérons. Cette post-stratification devrait permettre de réduire la variance et les effets de distorsion, mais cela variera selon notre compétence en matière de définition des cas de l'imputation; c'est là un domaine où il reste encore beaucoup de travail à abattre. *Estimation par la méthode de régression ou par la méthode du quotient* – Nous songeons aussi à effectuer l'estimation à l'intérieur des cas par la méthode de régression ou du quotient au lieu de la méthode hot deck. Par exemple, $\hat{z} = x_i - f x_j$, où $f = y/x$ serait calculé

Tableau A
Caractéristiques choisies des divers estimateurs

Estimateur	Biais	Variance	Réussite au contrôle de validation?
Echantillon complet	0	$\text{Var}(\bar{z})$	Oui
Double échantillon	0	$\text{Var}(\bar{z}) + c_1 S_B^2(Y)$	Oui
Méthode "hot deck"	0 ^a	$\text{Var}(\bar{z}) + c_1(1 + c_2) S_B^2(Y)$	Non
Rapport (R)	b ₁	$\text{Var}(\bar{z}) + V_1$	Oui
Par le quotient combiné	b ₂	$\text{Var}(\bar{z}) + V_2$	Oui

Note: En général, on peut dire que la procédure "hot deck" de base n'est pas biaisée uniquement lorsqu'elle produit des valeurs finales qui réussissent à passer tous les contrôles de validation.

Dans le tableau A, nous utilisons les caractéristiques de Z comme point de repère pour comparer entre eux les divers estimateurs.

Echantillonnage à deux degrés - En utilisant la notation de Cochran (Cochran 1977, 12.2), on a maintenant stratifié l'échantillon original de dimension n' en deux groupes, A et B, renfermant respectivement n_A' et n_B' unités. On extrait ensuite un sous-échantillon de dimension n_B du groupe B. Le montant original inscrit par le contribuable (X) est enregistré pour chacun des articles $n' = n_A' + n_B'$. Les changements dus au contrôle de l'élément "Autres revenus" (Y) sont enregistrés pour les n_A' unités du groupe A et pour le sous-échantillon aléatoire de n_B unités provenant du groupe B.

Puisque la procédure de double échantillonnage s'applique uniquement à la variable Y à l'intérieur du groupe B, l'estimateur de double échantillonnage de Z est

$$\bar{z}_d = \bar{x} - \bar{y}_d$$

$$= \bar{x} - (\sum y_{Ai} + (n_B'/n_B) \sum y_{Bj}) / n'$$

et \bar{z}_d est non biaisé.

Disons que N_B = nombre d'unités de population dans la strate B

$P_B = N_B/N$, proportion de population dans la strate B

\bar{Y}_B = moyenne de la population dans la strate B

$S_B^2(Y) = \Sigma (Y_{Bi} - \bar{Y}_B)^2 / (N_B - 1)$, $i = 1, 2, \dots, N_B$

$1/K$ = proportion du sous-échantillonnage, c'est-à-dire n_B/n_B' .

Si on suppose que la proportion d'échantillonnage $1/K$ est fixe (dans notre application, $1/K = .10$ ou $.20$), il s'ensuit que la variance inconditionnelle de \bar{z}_d (Cochran 1977) est

$$\text{Var}(\bar{z}_d) = \text{Var}(\bar{z}) + c_1 S_B^2(Y),$$

$$= [S^2(Z) + P_B(K - 1)S_B^2(Y)]/n',$$

où $c_1 = P_B(K - 1)/n'$.

REMERCIEMENTS

Les auteurs du présent document aimeraient remercier les membres du personnel de la Statistics of Income Division pour l'aide considérable qu'ils leur ont apportée; leurs responsabilités quotidiennes sont traitées dans le présent document. Nous aimerions aussi remercier David W. Chapman et John L. Czajka pour leurs nombreuses remarques constructives et bien entendu, toute responsabilité quant aux points obscurs ou erreurs qui pourraient encore s'y trouver.

ANNEXE: THÉORIE ÉLÉMENTAIRE

On décrit, dans la présente annexe, certains détails techniques relativement à la procédure d'échantillonnage à deux degrés appliquée à notre situation précise. Nous comparons plusieurs estimateurs éventuels pour le plan de double échantillonnage que nous avons choisi. On trouve au tableau A de cette annexe un résumé global des distorsions et variances pour les différentes approches.

Pour les fins de la présentation, nous n'avons pas tenu compte du plan d'échantillonnage stratifié sous-jacent et nous agissons comme si nous avions fait un simple échantillonnage aléatoire ou, ce qui est équivalent, nous tenons compte des estimations à l'intérieur d'une seule strate d'échantillonnage. Toute autre façon de procéder rendrait excessivement complexe la notation, mais ne modifierait en rien les principaux points que nous désirons exposer. Prenons donc un seul des éléments soumis au sous-échantillonnage, soit "Autres revenus" comme auparavant. La variable intéressante est Z , c'est-à-dire la valeur corrigée finale pour Autres revenus, et celle-ci peut se décomposer de la façon suivante:

$$Z = X - Y,$$

où X = valeur originale inscrite par le contribuable ou traitée par le service fiscal pour l'élément "Autres revenus"

Y = modification apportée à l'élément "Autres revenus" après la révision de l'annexe. Les valeurs des populations et les paramètres sont représentés par des majuscules et les statistiques de l'échantillonnage par des minuscules. Les paramètres d'intérêt relativement à la population sont la moyenne et la variance pour la population finie:

$$\bar{Z} = \sum Z_i / N = \bar{X} - \bar{Y}, \text{ et}$$

$$S^2(Z) = \sum (Z_i - \bar{Z})^2 / (N - 1).$$

Echantillon complet - Avant la mise en oeuvre du double échantillonnage, on calculait les estimations à partir d'un échantillon complet de dimension n' et l'estimateur sans biais de Z était:

$$\bar{z} = \sum z_i / n'$$

$$= \bar{x} - \bar{y}.$$

Si on ne tient pas compte de la correction de la population finie (N est très grand), la variance est:

$$\text{Var}(\bar{z}) = S^2(Z) / n'.$$

- Enfin, nous aimerions aussi, d'une certaine façon, établir nos estimations à partir des données de l'année précédente afin de pouvoir imputer les renseignements manquants plus tôt au cours du procédé. Pour minimiser le regroupement des cas d'ajustement, en 1981 et 1982 nous avons dû attendre que tous les articles soient disponibles avant d'effectuer l'imputation. Cette contrainte a retardé la production de plusieurs semaines. Nous pourrions régler ce problème en augmentant encore plus le nombre de donneurs, mais le contrôle d'un plus grand nombre d'articles comporte un inconvénient évident, c'est-à-dire l'augmentation des coûts. D'autre part, en fondant notre approche partiellement sur les données de l'année précédente on pourrait non seulement améliorer l'estimation, mais aussi faire les calculs des imputations durant l'exécution du traitement principal.

Résumé

Nous avons, dans le présent document, décrit les raisons pour lesquelles nous devons apporter des modifications importantes à notre traitement statistique des déclarations d'impôt provenant de sociétés:

- On a rejeté la méthode traditionnelle, c'est-à-dire l'estimation à partir des données complètes, en faveur de l'échantillonnage à deux degrés à cause de considérations financières.
- On a rejeté l'estimation habituelle dans le contexte de l'échantillonnage à deux degrés, c'est-à-dire la pondération des données complètes, parce qu'elle ne produit pas un ensemble de données rectangulaire.
- On a rejeté l'approche "hot deck" traditionnelle parce que les estimations obtenues grâce à cette méthode ne subsistaient pas toujours avec succès les contrôles de validation.

En lieu et place de ces méthodes, nous avons choisi d'estimer les changements relatifs grâce à l'imputation "hot deck" avec rapport à l'intérieur des cases d'ajustement. Nous avons supposé que, la procédure d'échantillonnage à deux degrés étant restreinte à un sous-ensemble de petites sociétés, les estimations qui intéressaient nos principaux utilisateurs ne subiraient presque pas d'effet défavorable; en réalité, ces estimations pourraient être améliorées grâce à une meilleure répartition des ressources qui nous permettrait de mieux valider et corriger les dossiers des grandes sociétés. Les résultats obtenus jusqu'à ce jour corroborent largement ces hypothèses.

Comparativement à l'estimation traditionnelle faite à partir de données complètes, le double échantillonnage et l'imputation "hot deck" font augmenter l'erreur quadratique moyenne des estimations de deux façons: d'abord par l'introduction d'une distorsion et ensuite par l'augmentation de la variance de l'estimation. Nos résultats préliminaires indiquent que l'effet de distorsion pourrait être important pour certaines estimations; toutefois, les exemples utilisés avaient d'abord été choisis parce qu'ils constituaient des cas où la méthode hot deck serait la moins efficace. Malgré tout, l'effet global évalué de cette procédure sur l'erreur quadratique moyenne semble relativement petit. Si on analyse l'augmentation de la variance, la portion la plus importante de cette augmentation est habituellement due à la dimension réduite de l'échantillon (double échantillonnage). Cette augmentation de la variance s'est aussi révélée relativement petite puisqu'un seul élément du montant final (la modification) est imputé; la variance des valeurs originales semble dominer la variance des modifications.

En conclusion, bien qu'il faille encore apporter plusieurs modifications à cette méthode, les résultats obtenus nous incitent à poursuivre l'utilisation de la technique actuelle, c'est-à-dire plan d'échantillonnage à deux degrés et imputation. Peut-être pourrions-nous, à un autre colloque de ce genre, présenter des résultats plus poussés de cette recherche.

Tableau 2

Distorsion relative approximative pour Recettes d'entreprises et Autres revenus, par industrie mineure choisie, 1982

Industries mineures choisies	Recettes d'entreprises		Autres revenus	
	Toutes Actifs	déclats inférieurs à \$25 millions	Toutes Actifs	déclats inférieurs à \$25 millions

(Les distorsions sont exprimées en pourcentage du total applicable)

COMMERCE EN GROS

Machinerie, pièces d'équipement et fournitures

COMMERCE DE DÉTAIL

Vendeurs d'automobiles et stations-service

FINANCE ET ASSURANCE

Institutions bancaires
Organismes de crédit sauf les banques

Courtiers d'assurance

Note: Tous les calculs sont faits à partir d'estimations des distorsions pondérées en fonction du plan. Les industries ont été choisies parce qu'elles représentaient les pires cas possibles.

Nous n'avons pas fait appel à l'échantillonnage à deux degrés et l'imputation pour les échantillons de 1983 et 1984 à cause de contraintes de traitement. On utilisera de nouveaux échantillons de 1985. Au moment de rétablir le procédé d'imputation, nous prévoyons y apporter de nombreux changements:

- Il ne sera plus nécessaire de transcrire certains éléments à des fins statistiques avant de soumettre les articles au double échantillonnage. On obtient maintenant directement du système de traitement des revenus de l'IRS les zones nécessaires de sorte que celles-ci sont disponibles avant même que nous ne commençons la lecture et le contrôle des déclarations d'impôt. Nous pouvons donc, avant que les préposés au contrôle ne fassent l'analyse primaire d'une déclaration, déterminer quelles sont les annexes qu'ils devraient examiner. Cette caractéristique ajoute encore à l'attrait du double échantillonnage stratifié car les économies devraient augmenter.

- Toutefois, à cause du nouveau système de traitement, nous ne pouvons maintenant appliquer le sous-échantillonnage qu'à trois annexes. Pour 1985, il s'agit des annexes "Autres revenus", "Autres déductions" et "Autres coûts des biens vendus"; les quatre autres annexes utilisées en 1981 et 1982 ont dû être retirées du plan de sous-échantillonnage. Malgré le succès modeste des procédures employées en 1981 et 1982, on modifiera les méthodes d'imputation pour 1985. Par exemple, on pourrait améliorer la définition courante des cases d'ajustement et il faut reconsidérer l'imputation séparée variant selon le modèle des éléments représentés. Il serait aussi souhaitable d'analyser la possibilité d'utiliser l'appariement prévisionnel des moyennes à l'intérieur des cases d'ajustement (Litle 1986). Nous devrions aussi, pour 1986, songer à perfectionner le plan de sous-échantillonnage.

et le montant original, est toujours petit ou si R est constant à l'intérieur des cases d'ajustement. Nous avons choisi d'analyser les pires cas possibles de distorsion en étudiant certains exemples où R n'est ni petit ni constant. Nous avons concentré notre attention sur deux variables: autres revenus et recettes d'entreprises.

Modèle non biaisé

La distorsion causée par le rapport dans l'imputation par la méthode hot deck que nous utilisons serait égale à zéro si la relation $Y = RX$ était vraie pour tous les membres de chaque case d'ajustement choisie. Il serait peut-être utile d'avoir un graphique global des données pour pouvoir examiner dans quelle mesure ce modèle se confirme pour l'élément "Autres revenus". Comme on peut le voir à la figure 2, nous avons par conséquent représenté graphiquement les donneurs du groupe B en séparant les sociétés financières des sociétés non financières. Il existe une différence certaine entre ces deux catégories. Les déclarations provenant de sociétés non financières sont beaucoup moins susceptibles d'être modifiées; en 1982, 14% des donneurs non financiers ont subi une modification à l'élément "Autres revenus" comparativement à 59% dans le cas des articles provenant de sociétés financières. En outre, il semble que le modèle $E(Y) = RX$ soit approprié, du moins dans le cas des déclarations provenant de sociétés financières. Nous avons l'intention de poursuivre le travail à cet égard, mais le pointage de dispersion nous encourage à croire que les distorsions existantes seraient pour une grande part petites.

Mesure des distorsions réelles

Le tableau 2 montre les distorsions relatives pour certaines industries choisies qui correspondent aux pires cas possibles. Les nombres indiqués correspondent à toutes les déclarations de l'industrie et aux déclarations où l'actif est inférieur à \$25 millions, c'est-à-dire celles des sociétés qui devraient selon toute vraisemblance être le plus touchées par la mise en oeuvre des nouvelles procédures. De tous les éléments modifiés par l'échantillonnage à deux degrés, l'annexe "Autres revenus" est celle pour laquelle les valeurs de R sont les plus grandes et les distributions de R les plus dispersées. Au contrôle de l'élément "Autres revenus" le plus grand changement concernait le montant des recettes d'entreprises. Il faut remarquer que les estimations de distorsions décrites au tableau 2 sont sujettes à une erreur d'échantillonnage considérable (Czajka 1986). Toutefois, à l'exception des plus petits montants, nous croyons que les estimations indiquées portent probablement le bon signe et correspondent à un ordre de grandeur approprié.

Ces exemples indiquent qu'à l'intérieur des petites sous-populations les effets de distorsion peuvent être importants. Toutefois, même dans une grande industrie, choisie précisément à cause des problèmes possibles, la distorsion pour les entreprises de toutes dimensions demeure relativement petite. Les résultats obtenus par Czajka (1986) indiquent que dans les estimations globales pour toutes les industries la distorsion causée par l'imputation demeure petite, c'est-à-dire inférieure à 1% dans tous les cas et considérablement moindre que .05% dans la plupart des cas.

Il ne fait aucun doute que certaines distorsions mentionnées au tableau 2 semblent importantes et justifient l'inquiétude; toutefois, il faut réaliser que l'effet global de l'écart quadratiques moyen de la distorsion est minime pour toutes les déclarations, il se situe généralement à 5% ou moins. Ces résultats prouvent suffisamment que les procédures employées ont peu ou pas nui aux données requises par nos utilisateurs; cela ne signifie pas toutefois que des améliorations importantes comme celles prévues pour 1985 et 1986 ne devraient pas être mises en oeuvre.

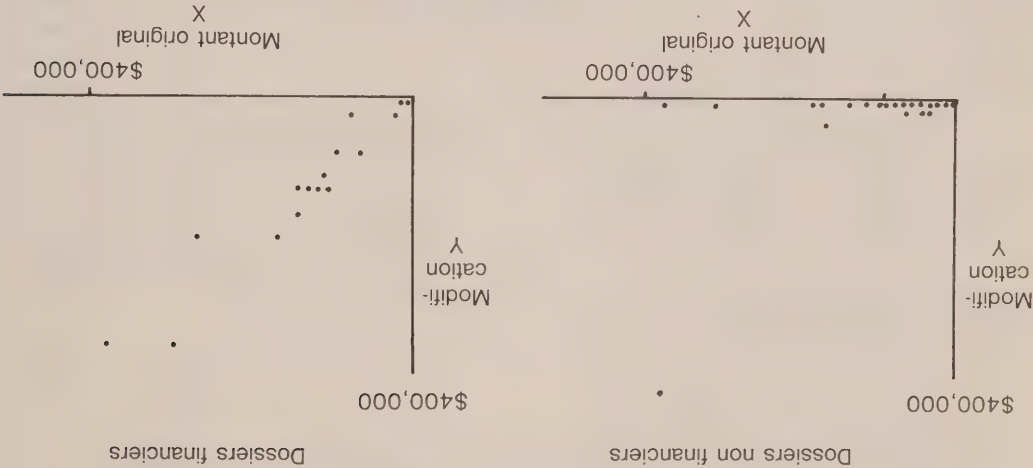


Figure 2. Modifications apportées à l'élément "Autres revenus": donneurs du groupe B seulement

Forts de notre expérience de 1981, nous avons apporté plusieurs modifications au plan d'échantillonnage à deux degrés de 1982:

- Etant donné l'importance du regroupement des cas dans le cas des déclarations d'institutions financières en 1981, le taux de sous-échantillonnage pour les déclarations de petites entreprises financières a été doublé en 1982, passant de 10% à 20% comme nous l'avons mentionné antérieurement; cette opération devait améliorer les estimations.
- En 1981, la procédure d'échantillonnage à deux degrés n'avait pas été appliquée à l'échantillon complet, mais avait plutôt été restreinte à certains centres de traitement. Les autres centres de traitement recueillaient tous les renseignements, comme auparavant. En 1982, la procédure a été appliquée à l'échantillon complet. En 1982 le nombre relatif d'articles dont certains éléments avaient été imputés est passé à 63% comparativement à 40% en 1981.
- Afin d'évaluer approximativement la variance des imputations par la méthode hot deck (Oh et Scheuren 1980; Rubin et Schenker 1986), on a imposé une restriction additionnelle au plan de 1982 en exigeant qu'il y ait au moins deux donneurs pour chaque case d'ajustement (voir le tableau 1).

En 1982, 54,196 articles devaient être imputés à partir de 6,503 donneurs et la quantité de regroupements des cas d'ajustement fut considérablement moindre (Hinkins 1984). Plus précisément, dans le cas des articles financiers, 94% des articles imputés en 1982 faisaient partie de cas d'ajustement définies grâce à certaines distinctions de dimension, comparativement à 75% en 1981. Le tableau 1 fournit certaines autres statistiques relative au fonctionnement des systèmes en 1981 et 1982.

4. EVALUATION INITIALE DE LA DISTORSION

L'évaluation du système d'échantillonnage à deux degrés de 1982 est en cours, mais nous avons déjà certains résultats quant aux effets possibles de distorsion de l'imputation. La distorsion devrait être faible si R, c'est-à-dire le rapport entre la modification apportée au contrôle

Tableau 1
Statistiques choisies sur la méthode d'imputation "hot deck" avec rapport, 1981-1982

Élément	Année d'imposition		Année d'imposition	
	Non	Financier	Financier	Non
NOMBRE				
Donneurs	908	3,081	1,806	4,697
Imputations	7,912	28,674	10,719	43,477
Cases d'ajustement	113	238	142	260
DIMENSION DES CASSES "DONNEUR"				
Moyenne	8	13	13	18
Maximum	68	58	126	98
Minimum	1	1	2	2
RAPPORT ENTRE DONNEUR ET IMPUTATION				
Moyen	.11	.11	.17	.11
Maximum	1.00	.25	2.00	.28
Minimum	.05	.05	.05	.05

Note: Pour 1982, il fallait absolument utiliser des cases de 2 donneurs pour qu'il soit possible de calculer la variance.

les cases d'ajustement sont définies en termes de classification industrielle, dimension de l'entreprise et modèle des éléments inclus dans la déclaration. Trente catégories ont été définies grâce aux divers critères de classification industrielle et de dimension (voir figure 1). De plus, seize modèles d'éléments ont été traités séparément selon la présence ou l'absence des articles Autres revenus (2 catégories), Autres déductions ou Autres coûts des biens vendus (2 catégories), Autre actif à court terme ou Autre actif à long terme (2 catégories) et finalement Autre passif à court terme ou Autre passif à long terme (2 catégories). Le nombre maximum de cases d'ajustement était donc $30 \times 16 = 480$.

On a élaboré une structure hiérarchique pour tous les modèles d'éléments afin de pouvoir faire un regroupement lorsque le nombre de donneurs servant à l'imputation était insuffisant (voir figure 1). La première division sépare les déclarations financières (banques, compagnies d'assurance, etc.) des déclarations de sociétés non financières; les cases ne sont pas regroupées pour cette division. Aux niveaux suivants de la hiérarchie, on sépare les cases en fonction de catégories industrielles relativement vastes et de la dimension des sociétés en termes d'actif et de bénéfice net. Il faut se rappeler que les très grandes entreprises ne font pas l'objet d'un sous-échantillonnage et que, par conséquent, elles ne devraient pas nécessiter d'imputation; il semblait donc suffisant d'établir des groupes en fonction de la catégorie industrielle et de la dimension.

La qualité de notre estimation dépend de la quantité de regroupements. En 1981, nous avions 36,586 déclarations avec au moins une annexe à imputer et 3,989 donneurs. Dans le cas des déclarations non financières, nous n'avons jamais fait de regroupement en fonction des principales classifications industrielles et, en fait, nous avons toujours établi une distinction en fonction de la dimension. De nombreuses cases n'ont jamais été combinées et au contraire on y a conservé le maximum de détails possibles. Dans le cas des déclarations de sociétés financières par ailleurs, la variable dimension a souvent été perdue lors du regroupement financier de toutes les cases et il est arrivé que des industries importantes soient combinées (Hinkins 1983). Dans un modèle, toute les déclarations de sociétés financières ont été regroupées dans la même case.

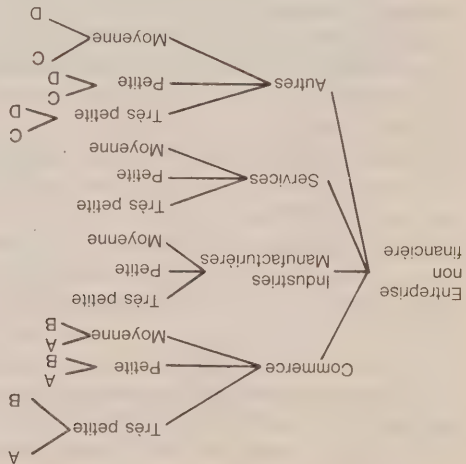


Figure 1. Hiérarchie des cas de d'ajustement pour la méthode "hot deck"

Nous avons prévu que grâce à ce calcul d'un rapport, en plus de satisfait aux contrôles de validation, on pourrait réduire la variance des estimations comparativement à celle obtenue grâce à l'approche hot deck de base; toutefois, la variance d'une estimation ne peut être établie par analyse; elle doit être mesurée empiriquement. Nous n'avons pas encore vérifié, dans notre application de cette méthode aux sociétés, la variance plus petite que nous prévoyions; toutefois, les résultats des simulations appuient l'approche choisie. En introduisant ainsi un rapport, nous obtenons cependant des estimations biaisées. Nous avons supposé que les distorsions seraient faibles et elles l'ont été en réalité, dans la majorité des cas, comme nous le montrons plus loin.

Le modèle correspondant à notre procédure d'imputation est fondé sur la définition des strates dans l'échantillonnage à deux degrés et des cas de d'ajustement. Nous avons pris plusieurs mesures constructives afin que cette approche soit raisonnable. Lors de la stratification initiale, nous nous sommes efforcés d'inclure dans le sous-échantillon uniquement les articles où les modifications seraient vraisemblablement petites ou nulles. En outre, nous avons *subjectivement* choisi les cas de d'ajustement afin qu'ils soient homogènes quant à l'ampleur des modifications relatives qui seraient apportées à la suite du contrôle. Plus précisément,

$$z_i = x_i - \bar{y} = 10,000 - (1/2)10,000 = +5,000.$$

Puisque le montant des autres revenus ne doit pas être négatif, les contrôles de validation signaleraient une erreur et des modifications additionnelles devraient être apportées à l'article en question. (Voir Sande 1982 pour obtenir une explication générale de ce problème). Puisque nous observons toujours le montant original, il semble plus raisonnable de procéder par la méthode "hot deck" pour établir la modification relative $R = Y/X$ au lieu d'adopter la modification réelle Y . Pour suivre l'exemple, puisque dans l'article donneur on supprime la moitié du montant inscrit dans "Autres revenus" à la lecture de l'annexe, il faut aussi supprimer la moitié de cette somme dans l'article imputé. Le montant final estimé pour les autres revenus est alors:

le groupe A les cas où les autres revenus dépassent une certaine valeur en dollars. Malheureusement, cette opération ne s'est faite qu'indirectement. Par inférence, on suppose que le groupe B inclut seulement les déclarations des petites entreprises pour lesquelles nous croyons que le contrôle donne lieu à peu ou pas de modification. (Pour d'autres détails à ce sujet, voir Barker *et coll.*, 1982.)

Dans le cas des déclarations cruciales faisant partie du groupe A, nous observons entièrement toutes les variables, c'est-à-dire tous les éléments. Seules les déclarations du groupe B sont visées par le sous-échantillonnage des sept annexes mentionnées plus haut. Même dans le cas des déclarations du groupe B, les montants originaux de tous les éléments sont toujours notés; par conséquent, on obtient certains renseignements au sujet de chaque élément. Les seuls renseignements qu'on n'obtient pas pour certains éléments du groupe B sont les modifications résultant du contrôle d'une annexe. C'est ces modifications qu'on fait en utilisant la procédure d'imputation qui est décrite dans la prochaine section. Les variables ne sont pas toutes touchées par le sous-échantillonnage; par exemple, des 600 éléments repérés dans le programme des déclarations des sociétés de 1981, seuls 56 furent touchés d'une façon quelconque par l'échantillonnage à deux degrés. Toutefois, des quelques 100 principaux éléments qui forment le revenu et le bilan, approximativement la moitié pourraient être touchés.

3. LA PROCÉDURE D'IMPUTATION

Les renseignements manquants (c'est-à-dire, les modifications résultant du contrôle) dans le groupe B sont imputés grâce à une procédure "hot deck" appliquée à l'intérieur des cases d'ajustement. Grâce à cette technique, on fait correspondre un article dont les annexes doivent être imputées à un article donneur dont les mêmes annexes sont contrôlées, dans la même case d'ajustement. On décrit la formation des cases d'ajustement dans cette section. En 1981, le taux de sous-échantillonnage était de 10% pour les déclarations soumises au sous-échantillonnage; une déclaration sur dix était choisie systématiquement en vue du contrôle (donneurs pour la procédure hot deck) et les neuf autres devaient être imputées. En 1982, on a maintenu le taux de sous-échantillonnage à 10% pour les déclarations non financières (commerce, fabrication, etc.), mais on l'a fixé à 20% dans le cas des déclarations financières (banques, compagnies d'assurance, etc.).

Dans une case d'ajustement, le nombre de déclarations n' peut être divisé en nombre de donneurs n'' et nombre d'éléments imputés $n' - n''$. Étant donné le taux réduit de sous-échantillonnage, le nombre de donneurs est presque toujours inférieur au nombre d'éléments imputés. Prenons l'exemple où $n' = n'' + t$ où t est le nombre de donneurs entiers non négatifs et où $0 \leq t < n''$. Dans un tel cas, la procédure hot deck choisit les $n' - n''$ donneurs t fois et choisit les t unités qui restent par simple échantillonnage aléatoire, sans remplacement. Pour poursuivre notre illustration, rappelons-nous que l'élément qui nous intéresse est Z , c'est-à-dire le montant final corrigé des autres revenus. Z peut prendre la forme $Z = X - Y$ où X représente le montant original inscrit par le contribuable dans Autres revenus et Y représente la modification apportée à la suite du contrôle de l'annexe "Autres revenus". Seule la modification Y n'est pas observée et doit être calculée approximativement pour un sous-ensemble de déclarations faisant partie du groupe B.

Si nous employons simplement la procédure hot deck traditionnelle et que nous estimons la valeur y_i non observée pour l'article i , lorsque la valeur observée pour l'article donneur j est y_j , l'estimation obtenue pour la valeur finale z_i peut ne pas satisfaire aux contrôles de validation. Par exemple, disons que dans l'article donneur la valeur inscrite à l'origine dans la section Autres revenus était de \$30,000 et qu'on a soustrait \$15,000 de cette valeur au moment du contrôle. Supposons que dans l'article à imputer le montant original des autres revenus est égal à \$10,000, la modification imputée de \$15,000 produirait alors une estimation négative des autres revenus:

$$z_i = x_i - y_i = 10,000 - 15,000 = -5,000.$$

Pièce 2

Illustration du contrôle de l'élément "Autres revenus"

Genre de revenu	Montant original(\$)	Montant de la modification(\$)	Montant final(\$)
Autres revenus	1,600	- 1,200	400
Recettes	500	+ 900	1,400
Loyers	0	+ 300	300
Intérêts	700	0	700

- Décider de revoir ou non une annexe en particulier à partir des renseignements originaux transcrits. De nouveau, avant le programme de 1981, les préposés au contrôle devaient bien entendu contrôler entièrement toutes les annexes.

Pour l'analyse des déclarations de sociétés de 1981 et 1982, nous avons choisi sept éléments pour le sous-échantillonnage, de même que les annexes correspondant à ceux-ci: autres revenus, autres déductions, autres coûts des biens vendus, autre actif à court terme, autre actif à long terme, autre passif à court terme et autre passif à long terme.

On peut changer substantiellement les montants provenant des déclarations d'impôt des sociétés lors du contrôle. Par exemple, prenons l'annexe "Autres revenus" illustrée à la pièce 2. Les sommes originales (colonne 1) sont initialement observées pour chaque déclaration. Les variables qui font l'objet du sous-échantillonnage sont les modifications qui seraient apportées si l'annexe "Autres revenus" était contrôlée (colonne 2). Dans le cas hypothétique présenté ici le montant original des autres revenus est \$1,600; une fois examiné par le préposé, ce montant serait reclassifié et réparti comme suit: \$900 - Recettes d'entreprise, \$300 - Loyers, \$400 qui appartiennent réellement à la catégorie Autres revenus. Les variables intéressantes sont bien sûr les montants corrigés correspondant à chaque élément.

Avant de mettre en oeuvre le nouveau système de traitement, nous avons fait une expérience pour comparer la durée de la transcription initiale avec contrôle réduit à la durée du contrôle complet avec lecture de toutes les annexes. Comme nous l'avions prévu, le contrôle réduit se faisait beaucoup plus rapidement et, par conséquent, à un moindre coût. Il était donc possible d'économiser beaucoup de ressources en effectuant un sous-échantillonnage. Nous avons, de manière très conservatrice, extrapolé les économies de coût pour 1981 à \$300,000 au moins, en supposant qu'on faisait partiellement appel à la technique du sous-échantillonnage.

Double-échantillonnage

Nous sommes maintenant prêts à décrire la stratification bidimensionnelle de base adoptée pour notre échantillonnage à deux degrés. Les déclarations sont classées en deux groupes: déclarations cruciales (groupe A) et autres (groupe B). Les déclarations cruciales regroupent toutes celles des sociétés dont l'actif est égal ou supérieur à \$50 millions; ce groupe inclut donc toutes les déclarations de très grandes entreprises, la plupart des déclarations choisies avec certitude pour faire partie de l'échantillon et les déclarations de toutes les sociétés, quelle que soit leur dimension, lorsque la probabilité de correction lors du contrôle est élevée. Nous désirons de toute évidence obtenir un plan de sous-échantillonnage grâce auquel on pourrait contrôler toutes les annexes où la probabilité de modification était élevée, particulièrement s'il s'agissait d'une modification importante, et faire un sous-échantillonnage des autres. Pour tenter de prévoir quelles annexes seront probablement modifiées, on inclut un article dans le groupe A si le montant original indiqué dans Autres revenus, pour poursuivre avec le même exemple, est anormalement élevé comparativement au montant du revenu total. En outre, puisque nous ne désirons pas imputer de très gros montants, il faut aussi inclure dans

L'extraction des informations de chacune des déclarations constituant l'échantillon est un procédé onéreux et laborieux. On peut extraire au-delà de 600 éléments d'une déclaration et ceux-ci ne sont pas simplement extraits, ils doivent aussi être vérifiés attentivement et redistribués pour compenser les variations dans les façons de remplir les déclarations. Le procédé complet s'intitule "contrôle d'une déclaration" et le coût de ce "contrôle" varie selon le degré de complexité. Il suffit parfois de vingt-cinq minutes pour contrôler une déclaration relativement simple, mais dans le cas d'une déclaration réellement complexe, le procédé peut durer jusqu'à une semaine. La qualité de ce contrôle est vitale pour les estimations car les vérifications effectuées réduisent les inconsistances rencontrées, mais ne les éliminent pas entièrement.

En fait les erreurs non dues à l'échantillonnage constituent l'une des embûches importantes du procédé de contrôle des données, particulièrement dans le cas des très grandes sociétés. Afin de consacrer proportionnellement plus de ressources à la diminution des erreurs non dues à l'échantillonnage pour les déclarations des grandes sociétés, nous avons mis en place l'échantillonnage à deux degrés stratifié pour les déclarations des petites entreprises; plus précisément, certains éléments de données sont extraits uniquement d'un sous-échantillon de déclarations, c'est-à-dire un sous-ensemble de déclarations où les actifs sont inférieurs à \$50 millions. Même si cette modification pouvait donner lieu à l'augmentation des erreurs dans le cas de certaines variables des déclarations des petites entreprises, nous prévoyions que ces procédures auraient peu d'effets négatifs sur les estimations des totaux à l'échelle du pays ou sur les estimations des sous-domaines d'intérêt primordial pour nos principaux utilisateurs. Deux raisons majeures justifiaient cette conjecture:

- Comme nous l'avons déjà mentionné, les déclarations des sociétés dont l'actif total est égal ou supérieur à \$50 millions n'ont pas été soumises à cette étape additionnelle d'échantillonnage.
 - La perte d'information due au sous-échantillonnage a été réduite grâce au choix des éléments ou variables visés par ce sous-échantillonnage.
- Comme nous le montrons par la suite, les résultats obtenus jusqu'à ce jour confirment amplement nos prévisions.

Éléments choisis pour le sous-échantillonnage

Lorsque certains éléments divers d'une déclaration sont différents de zéro, le contribuable doit joindre une annexe pour fournir des renseignements additionnels. Par exemple, si la case "Autres revenus" renferme une valeur différente de zéro, la société doit décrire l'objet de cette valeur. Les annexes sont présentées sur des feuilles séparées dont le format et la longueur ne sont pas standard. Le procédé du contrôle d'une annexe se fait donc en diverses étapes: il faut d'abord trouver l'annexe, décider ensuite si le contribuable a indiqué des montants appropriés dans la section "Autres revenus" et faire les modifications qui s'imposent si on constate certaines erreurs.

À compter de l'année d'imposition 1981, pour le programme visant les sociétés, le contrôle de données statistiques provenant des déclarations d'impôt s'est faite par étapes et certains éléments ont été, au début, transcrits directement des déclarations pour être employés dans la compilation de statistiques. Grâce aux tests automatiques on pouvait par la suite marquer certains éléments ou annexes afin qu'ils fassent l'objet d'un dépouillement ou d'une analyse plus poussée lors des étapes subséquentes (Cys et coll. 1982). Grâce à cette nouvelle stratégie, nous avons pu:

- Conserver les renseignements originaux provenant des contribuables et évaluer ainsi l'importance des modifications faites au moment du contrôle. Avant l'échantillonnage de 1981, nous ne possédions aucun renseignement au sujet de l'ampleur des modifications apportées durant le contrôle. Les préposés au contrôle n'inscrivaient que le résultat final. (Voir Powell et Stubbs, 1981).

La technique d'imputation utilisée, c'est-à-dire l'imputation "hot deck", est simple. La nécessité d'expliquer la mise en application d'une procédure aussi simple pourra surprendre les théoriciens, mais, comme nous le montrons, les problèmes de mise en oeuvre dans le cadre d'une très vaste opération statistique sont nombreux.

Dans la suite du présent document, nous décrivons en détail la procédure d'échantillonnage à deux degrés et la technique d'imputation employée. Nous présentons aussi les résultats préliminaires obtenus à l'analyse de l'impact de ces procédures et nous décrivons, dans la dernière section, nos conclusions et projets futurs. Nous avons aussi ajouté en annexe une brève explication théorique des méthodes d'estimation employées et de leurs propriétés.

2. DESCRIPTION DES PROCÉDURES D'ÉCHANTILLONNAGE

L'IRS prélève un échantillon annuel des déclarations d'impôt des sociétés des États-Unis afin de faire une estimation des taux de variables économiques et fiscales pour tout le pays. En 1985 par exemple, environ 3 millions de déclarations d'impôt de sociétés seront présentées et l'échantillon de l'IRS sera formé de plus de 90,000 de ces déclarations. (Au Canada, il existe deux échantillons différents de déclarations d'impôt des sociétés, chacun étant conçu pour satisfaire à une gamme plus étroite de besoins. L'échantillon de Revenu Canada-impôt (voir Burpee et McGrath 1982) a été mis au point dans un but de simulation des politiques fiscales. L'échantillon de Statistique Canada (voir Ambrose 1985) vise principalement l'estimation d'aggrégats économiques. Nous croyons que nous pourrions améliorer l'efficacité des procédures actuelles aux États-Unis en mettant sur pied des plans séparés, mais des systèmes de traitement qui ne seraient pas entièrement distincts. Toutefois, le travail effectué jusqu'à ce jour (Clickner *et coll.* 1984) indique que les difficultés sont énormes et que le progrès se fait lentement.)

Les estimations annuelles ainsi obtenues se rapportent à la population complète des sociétés et à des sous-population habituellement définies par le genre d'industrie et la dimension des entreprises. La population sous-jacente est hautement asymétrique. Pour la plupart des variables, une très faible proportion de la population représente une fraction importante de la valeur totale en dollars. On trouve à la pièce 1 des exemples pour les déclarations des sociétés de 1982.

On utilise un plan d'échantillonnage hautement stratifié; la probabilité de sélection des petites sociétés est faible tandis que la sélection des grandes sociétés est presque certaine (Jones et McMahon, 1984). Les strates sont définies en fonction de la classification industrielle et de l'importance des sociétés en termes d'actif et de bénéfice net. Les probabilités de sélection pour chaque strate sont déterminées grâce à une forme modifiée de la répartition Neyman. Presque toutes les entreprises de la strate 100%, c'est-à-dire celles dont les déclarations sont choisies de façon certaine, ont un actif total de \$50 millions ou plus. On fait appel à une forme d'estimation par la méthode du quotient post-stratifiée pour pondérer les résultats de l'échantillon (Leszcz, Oh et Scheuren 1983).

Pièce 1

Degré de concentration des variables d'entreprises choisies

Eléments	Actif	Actif égal ou supérieur à \$50 millions
Nombre de déclarations	99,6%	0,4%
Total de l'actif	16,3	83,7
Total des recettes	39,3	60,7
Total de l'impôt	25,9	74,1

Source: Internal Revenue Service, 1985.

L'imputation par la méthode "hot deck" appliquée à un plan d'échantillonnage à deux degrés

SUSAN HINKINS et FRITZ SCHEUREN¹

RÉSUMÉ

À partir d'un échantillon annuel de déclarations d'impôt présentées par des sociétés américaines, le fisc américain (IRS) prépare des estimations de totaux de plusieurs centaines d'éléments financiers pour des ensembles et des sous-ensembles. Le plan d'échantillonnage de base est hautement stratifié et relativement complexe. Ce plan a été modifié lors de l'échantillonnage des déclarations de 1981 et 1982; on y a ajouté une procédure d'échantillonnage à deux degrés. Ce changement répondait à la nécessité d'une meilleure répartition des ressources dans un contexte de budgets en décroissance. Les éléments qu'on ne peut observer dans le sous-échantillon font l'objet d'une prédiction par imputation "hot deck" modifiée. Le présent document décrit la conception de cette nouvelle procédure de même que l'estimation et l'évaluation de ses effets.

MOTS CLÉS: Échantillonnage à deux degrés; hot deck; imputation.

1. INTRODUCTION

La première idée qui nous vient à l'esprit lorsqu'on entend parler du U.S. Internal Revenue Service n'est sans doute pas l'enquête par sondage. Toutefois, à chaque année au mois d'avril, les résidents des États-Unis prennent part à au moins une enquête administrative des qu'ils présentent une déclaration d'impôt individuelle. Nous faisons un échantillonnage annuel de ces données administratives à des fins statistiques. Un autre de nos principaux programmes vise l'échantillonnage annuel des déclarations d'impôt des sociétés américaines et c'est de cette enquête dont il sera question ici.

Ce qui nous intéresse en premier lieu lors d'un symposium comme celui-ci, ce sont les non-réponses ou les autres données manquantes. Malgré nos efforts importants en matière d'écution, nous connaissons aussi à l'IRS des problèmes de non-réponse. Toutefois, le présent document traite d'un genre différent de données manquantes, c'est-à-dire l'absence qui n'est pas imprévue, mais plutôt conçue (voir aussi Strudler, Oh et Scheuren 1986 qui donnent un autre exemple). Nous prenons la liberté d'analyser ces problèmes car nous utilisons pour les traiter une technique actuellement réservée aux non-réponses, c'est-à-dire l'imputation "hot deck" (Ford 1983). Le cas présente nous permet d'évaluer la procédure d'imputation puisque le mécanisme sous-jacent de non-réponse est bien connu.

Nous avons fait appel à l'échantillonnage à deux degrés pour analyser les déclarations d'impôt des sociétés afin de réduire les coûts tout en maintenant la perte d'information à un niveau tolérable. La ré pondération selon le degré d'échantillonnage constitue une approche d'estimation standard dans l'échantillonnage à deux degrés (voir par exemple Cochran 1977); toutefois, dans notre application il nous aurait fallu répondre presque tous les éléments un par un. Nos utilisateurs qui requièrent des ensembles de données rectangulaires ont jugé que cette méthode était inacceptable. On trouve la description d'une approche analogue utilisée dans un contexte canadien dans Colledge et coll. (1978).

¹ Susan Hinkins, Statistics of Income Division, Internal Revenue Service, boîte postale 369, Bozeman, Montana 59771. Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Avenue N.W., Washington, DC 20224.

- NELSON, D., McMILLEN, D., et KASPRZYK, D. (1985). An overview of the Survey of Income and Program Participation: Update I. SIPP Working Paper Series No. 8401, U.S. Bureau of the Census.
- OH, H.L., et SCHEUREN, F.J. (1980a). Estimating the variance impact of missing CPS income data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 408-415.
- OH, H.L., et SCHEUREN, F.J. (1980b). Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 416-420.
- OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys*, vol. 2, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 143-184.
- PALMER, S., et JONES, C. (1967). A look at alternate imputation procedures for CPS noninterview. *Proceedings of the Social Statistics Section, American Statistical Association*, 73-80.
- PAUL, E.C., et LAWES, M. (1982). Caractéristiques des ménages répondants et non-répondants dans l'enquête sur la population active. *Techniques d'enquête*, 8, 53-79.
- POLITZ, A., et SIMMONS, W. (1949). An attempt to get the 'Not-At-Homes' into the sample without callbacks. *Journal of the American Statistical Association*, 44, 9-31.
- ROSENBAUM, P., et RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- THOMSEN, I., et SIRLING, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. Dans *Incomplete Data in Sample Surveys*, Vol. 3, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 25-29.
- U.S. Department of Commerce, Bureau of the Census (1977). National Crime Survey, national sample, survey documentation. Rapport du Bureau of the Census.
- U.S. Department of Commerce, Bureau of the Census (1983). Cross-sectional weighting specifications for the first wave of the 1984 panel of the Survey of Income and Program Participation (SIPP). Note de service interne du Census Bureau adressée par C. Jones à T. Walsh, 25 novembre.
- U.S. Department of Commerce, Bureau of the Census (1984a). Economic characteristics of households in the United States: Third Quarter 1983. *Current Population Reports, Series P-70, No. 1*, Washington, D.C.: U.S. Government Printing Office.
- U.S. Department of Commerce, Bureau of the Census (1984b). 1984 SIPP first wave weighting-first stage estimate factors and specifications for collapsing noninterview adjustment calls. Note de service interne du Census Bureau adressée par C. Jones à T. Walsh, 16 février.
- U.S. Department of Commerce, Bureau of the Census (1984c). SIPP weighting: Subsequent wave cross-sectional - revised. Note de service interne de l'U.S. Bureau of the Census adressée par C. Jones à T. Walsh, 12 octobre.
- WELNIAK, E.J., et CODER, J.F. (1980). A measure of the bias in the March CPS earnings imputation scheme. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 421-425.
- YCAS, M., et LININGER, C. (1981). The income survey development program: Design features and initial findings. *Social Security Bulletin*, vol. 44, n° 11, novembre.

Division, Economic Surveys Division, Governments Division, Industry Division, Statistical Methods Division et Statistical Research Division.
Les auteurs désirent également remercier Dr. Fritz Scheuren pour ses commentaires pertinents lors de la rédaction de cet article.
Finalement, les auteurs remercient Hazel Beaton, Alice Bell et Valerie Howard pour leur excellent travail de dactylographie.

BIBLIOGRAPHIE

ANDERSON, H. (1978). On nonresponse bias and response probabilities, *Scandinavian Journal of Statistics*, 6, 107-112.

BAILEY, L. (1986). A study of alternative imputation techniques for surveys in the Current Industrial Reports. Rapport interne du Census Bureau, le 24 décembre.

BAILEY, L., CHAPMAN, D.W., et KASPRZYK, D. (1985). Nonresponse adjustment procedures at the Census Bureau: A review. *Proceedings of the Bureau of the Census First Annual Research Conference*, 241-444.

DAVID, M., LITTLE, R.J.A., SAMUEL, M.E., et TRIEST, R.K. (1986). Methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.

DYKE, T.C. (1984). Evaluation of the use of administrative record data for establishments which were non-respondents to the 1977 Census of Wholesale Trade, Retail Trade, or Selected Services. Rapport interne: Statistical Research Division Report Series, Census/SRD/RR-84/08, U.S. Bureau of the Census.

HANSON, R. (1978). The Current Population Survey: Design and Methodology. Technical Paper No. 40, Washington, D.C.: U.S. Bureau of the Census, pp. 55-59.

HUANG, E.T. (1986). Comparison of different imputation procedures in the monthly retail trade survey. *Proceedings of the Survey Research Methods Section, American Statistical Association* (à paraître).

KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.

KALTON, G., et LEPKOWSKI, J., et LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP research panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 372-377.

KALTON, G., McMILLEN, D. et KASPRZYK, D. (1986). Non-sampling error issues in the Survey of Income and Program Participation. *Proceedings of the Bureau of the Census Second Annual Research Conference*, 147-164.

KALTON, G., et MILLER, M. (1986). Effects of Adjustments for Wave Nonresponse on Panel Survey Estimates. *Proceedings of the Survey Research Methods Section, American Statistical Association* (à paraître).

KOBIŁARCZAK, E.L., et SINGH, R.P., (1986). SIPP: Longitudinal estimation for persons' characteristics. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (à paraître).

LILLARD, L., SMITH, J.P., et WELCH, F. (1982). What do we really know about wages: The importance of non-reporting and census imputation. *Journal of Political Economy*, 94, 489-506.

LITTLE, R.J.A. (1986). Missing data in Census Bureau surveys. *Proceedings of the Census Second Annual Research Conference*, 442-454.

LITTLE, R.J.A., et SAMUEL, M.E. (1983). Imputation models on the propensity to respond. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 415-420.

peuvent être utilisés à maintes reprises, ce qui augmente la variance. On pourrait modifier la méthode de façon qu'on ne puisse utiliser les valeurs des enregistrements donneurs plus d'une ou deux fois; cependant, cette modification n'a pas été apportée. La méthode "hot deck" utilisée pour la CPS repose sur l'hypothèse de la non-réponse sans biais, selon laquelle la distribution des réponses pour une variable d'enquête est la même chez les répondants et les non-répondants qui sont dans la même cellule.

David et coll. (1986) ont élaboré plusieurs méthodes basées sur des modèles, qui pourraient être utilisées à la place de la méthode "hot deck" hiérarchique employée pour la CPS. Ils les ont évaluées ainsi que la méthode "hot deck" hiérarchique par rapport à l'erreur absolue moyenne et à l'erreur relative moyenne.

Ces évaluations ont été faites à partir d'un fichier apparié CPS-IRS. Avec la création de ce fichier, on a tenté d'apparier le fichier de la CPS de mars 1981 aux dossiers fiscaux de l'IRS pour 1980. En dépit des limites apparentes de la méthode "hot deck", cette méthode appliquée à la CPS a permis d'obtenir une erreur absolue moyenne et une erreur relative moyenne plus faibles que lorsqu'on a utilisé des méthodes basées sur un modèle. Cependant, les modèles avaient été construits pour environ 10% seulement de tout l'échantillon de la CPS utilisé pour élaborer la méthode "hot deck".

6. SOMMAIRE ET CHAMPS D'ÉTUDE ÉVENTUELS

Dans cet article on a décrit, principalement à l'aide d'exemples, les méthodes de compensation de la non-réponse s'appliquant aux questionnaires utilisés actuellement dans les enquêtes et recensements de l'U.S. Bureau of the Census. On a mis l'emphasis sur le besoin d'effectuer d'autres études empiriques et théoriques tant dans le domaine démographique que dans le domaine économique afin d'établir des lignes directrices plus objectives pour a) concevoir des méthodes de compensation de la non-réponse et b) mesurer les effets de la non-réponse sur les résultats d'enquête, selon diverses conditions d'enquête. On a également mentionné quelques projets de recherche dans ce domaine serait fort souhaitable. Par exemple, l'augmentation de recherche dans ce domaine serait fort souhaitable. Par exemple, l'augmentation de recherche dans ce domaine serait fort souhaitable. Par exemple, l'augmentation de recherche dans ce domaine serait fort souhaitable.

Toutefois, l'augmentation de recherche dans ce domaine serait fort souhaitable. Par exemple, l'augmentation de recherche dans ce domaine serait fort souhaitable. Par exemple, l'augmentation de recherche dans ce domaine serait fort souhaitable.

En général, la compensation pour la non-réponse n'est qu'une étape parmi plusieurs permettant de réduire la variance et le biais des résultats des enquêtes et recensements. La réduction de l'impact de la non-réponse que pourraient apporter de telles étapes fera l'objet d'études ultérieures. De plus, il faut poursuivre les travaux de recherche concernant les questions qui reviennent périodiquement telles que l'effet des poids utilisés pour compenser la non-réponse au questionnaire et de l'imputation pour compenser la non-réponse à une question sur les estimateurs de la variance complexes, sur les méthodes de modélisation utilisées pour déterminer les facteurs de compensation appropriés ainsi que l'efficacité de la combinaison de divers types de techniques de compensation de la non-réponse.

REMERCIEMENTS

Les auteurs tiennent principalement à remercier tous ceux qui ont fourni des renseignements concernant les méthodes de compensation de la non-réponse utilisées dans beaucoup d'enquêtes et recensements de l'U.S. Census Bureau. Plusieurs divisions du Census Bureau ont contribué à cet article dont: Agriculture Division, Business Division, Construction Statistics

5. IMPUTATION POUR LES GAINS DANS LE CAS DE LA CURRENT POPULATION SURVEY

5.1 La méthode "hot deck" hiérarchique

La Current Population Survey (CPS) est une enquête permanente du Census Bureau qui vise environ 60,000 ménages américains chaque mois. La CPS, parrainée par le Bureau of Labor Statistics, cherche principalement à recueillir des renseignements sur la population active et sur l'emploi. Tous les ans, en mars, l'enquête de la CPS comporte un supplément sur le revenu. Entre 11 et 12% des ménages de l'échantillon ne répondent pas aux questions sur le revenu. Une procédure spéciale, la méthode "hot deck" hiérarchique, a donc été élaborée pour traiter les réponses manquantes par imputation.

Avec la méthode "hot deck" hiérarchique, les gains non déclarés sont pris en compte à partir de l'enregistrement des réponses d'une autre unité observée dans l'échantillon – un enregistrement donneur. Le but est de trouver un donneur dont les caractéristiques pour cette enquête ressemblent à celles du non-répondant à une question. La première étape dans la recherche des enregistrements donneurs appropriés consiste à séparer tout l'échantillon, à partir des cas de non-réponse totale au questionnaire, en cellules d'après une classification multiple basée sur un certain nombre de caractéristiques de l'enquête. Dans chaque cellule, on dresse une liste des répondants et des non-répondants à une question donnée. Des enregistrements donneurs tirés de la liste des répondants sont affectés systématiquement aux non-répondants, avec origine choisie au hasard. S'il y a plus de non-répondants que de répondants dans une cellule pour une question donnée, les réponses de certains ou peut-être de tous les répondants dans la cellule sont utilisées plus d'une fois. Il se peut que, dans certaines cellules, il y ait au moins un non-répondant mais aucun répondant à une question.

Pour éviter de se trouver avec des non-répondants sans enregistrements donneurs correspondants, la définition des cellules et le choix des enregistrements donneurs pour les non-répondants à une question sont effectués plusieurs fois. À chaque étape, on définit moins de cellules qu'à l'étape précédente. À l'étape finale, le nombre de cellules définies est suffisamment petit pour qu'on soit certain qu'il y aura des enregistrements donneurs dans chaque cellule. Pour définir les cellules à chaque étape, on combine les cellules utilisées à l'étape précédente. Chaque non-répondant à une question se voit donc attribuer au moins un enregistrement donneur. L'enregistrement donneur utilisé pour produire la valeur imputée est le premier qui est trouvé au cours des différentes étapes successives.

Le principal avantage que cette méthode hiérarchique offre, c'est qu'on peut définir un très grand nombre de cellules à la première étape, à cause de la série d'étapes successives utilisées. Chaque fois qu'on trouve un enregistrement donneur à la première étape, on applique le non-répondant à une question ainsi que le donneur en fonction d'un grand nombre de caractéristiques définies pour l'enquête. Dans de tels cas, il est fort probable qu'on puisse avoir une bonne imputation. Dans d'autres cas, les non-répondants à une question ainsi que les donneurs seront apparés en fonction d'un peu moins de caractéristiques. Cette méthode hiérarchique vise à trouver les enregistrements donneurs d'une façon qui maximise le nombre de caractéristiques pertinentes qui sont apparées. Pour une description plus détaillée de cette méthode, voir Welniak et Coder (1980), Oh et Scheuren (1980a), ou David, Little, Samuhel et Triest (1986, section 2).

5.2 Évaluation de la méthode "hot deck" hiérarchique

Certaines études portant sur la méthode "hot deck" hiérarchique utilisée pour la CPS ont été effectuées par Welniak et Coder (1980); Oh et Scheuren (1980a et 1980b); Lillard, Smith et Welch (1982); et David et coll. (1986). Une des faiblesses de la méthode "hot deck" employée pour la CPS qu'on a signalé est le fait que les valeurs des enregistrements donneurs

Pour le recensement de 1982, les non-répondants ont été classés comme "gros" ou "petits" selon que leurs ventes prévues étaient supérieures ou inférieures à \$100,000. Un suivi téléphonique-intégral a été effectué dans le cas de tous les gros non-répondants. Dans le cas des petits non-répondants, on les a stratifiés d'après d'autres caractéristiques de la liste d'envoi postal. On a fait le suivi par la poste et par téléphone auprès d'un échantillon de ces unités afin d'obtenir des estimations, par strate à l'intérieur de chaque État, du pourcentage de non-répondants qui sont effectivement des exploitations agricoles. Ces estimations ont ensuite été utilisées, avec les données sur les proportions de répondants visés par comté, pour produire des estimations du nombre d'exploitations agricoles non répondantes au niveau du comté pour chaque strate. Les poids d'un échantillon aléatoire de répondants par comté, qui était compatible avec le nombre estimé d'exploitations agricoles non répondantes, ont alors été multipliés par deux. Tous les autres répondants ont conservé leur poids unitaire.

4.4 Recherches relatives à la compensation de la non-réponse dans les enquêtes économiques

Les données de l'IRS tirées des déclarations d'impôt constituent probablement la source de renseignements la plus utile pour la compensation, par imputation, de la non-réponse au questionnaire dans les enquêtes économiques. Il peut exister certaines différences entre les données de l'IRS et celles recueillies dans les recensements économiques à cause des diverses définitions, formules et méthodes utilisées pour la collecte des données. Dans une étude (1984), Dyke a comparé les données administratives (IRS) utilisées pour l'imputation des ventes, des recettes, de la paye et de l'emploi dans le recensement du commerce de 1977 avec les réponses correspondantes obtenues pour un échantillon de suivi de non-répondants. Il a trouvé que, en général, les valeurs déclarées à l'enquête de suivi étaient supérieures à celles tirées des sources administratives. L'importance des écarts variait selon la question et, en outre, les écarts étaient plus prononcés dans le cas des entreprises à établissements multiples. Aussi, comparaisons additionnelles de ce genre s'imposent. Si des écarts systématiques sont notés, il sera alors possible d'élaborer des facteurs de compensation qui pourront être appliqués aux données de l'IRS.

Pour plusieurs recensements et enquêtes, un facteur "rapport d'éléments identiques" est calculé et appliqué aux chiffres d'une période antérieure afin de produire une valeur imputée pour la période courante. Il est possible que ce rapport calculé pour *tous* les cas dans l'échantillon qui ont répondu au questionnaire pour les deux périodes ne s'applique pas très bien aux non-répondants à certaines questions. Bailey (1986) a examiné l'utilisation d'autres moyens que le "rapport d'éléments identiques" pour imputer les valeurs manquantes, telles que la régression linéaire et la régression quadratique, à l'aide de divers ensembles de variables indépendantes.

Pour de nombreuses méthodes d'imputation pour la compensation de la non-réponse dans les enquêtes économiques, les cas dans l'échantillon, à la fois des répondants et des non-répondants, sont classés dans des cellules avant le calcul a) d'un certain type de rapport entre la période courante et la période antérieure, relativement à une question ou b) d'un certain type de relation entre les questions de l'enquête et les questions de base: paye, emploi et recettes. Un projet de recherche visant à étudier d'autres choix possibles pour la définition des cellules dans le cas de l'enquête mensuelle sur le commerce de détail vient d'être réalisé par Huang (1986). Elle a trouvé que, pour certains secteurs d'activité économique, une autre méthode de définition des cellules réduit considérablement l'erreur quadratique moyenne (EQM) des ventes estimées. De plus, elle a comparé la méthode d'imputation utilisée actuellement – les "rapports d'éléments identiques" – et trois autres méthodes, pour ce qui a trait au biais et à l'EQM. Une seule des méthodes évaluées a donné de meilleurs résultats que la méthode utilisée actuellement. Cependant, elle a conclu que les faibles avantages de la méthode optimale pourraient ne pas justifier les exigences additionnelles inhérentes à son application.

ce faire, on doit multiplier les ventes déclarées trois mois auparavant par le "rapport d'éléments identiques" basé sur la somme pondérée des ventes au cours du mois précédent et la somme pondérée des ventes trois mois auparavant (cellule par cellule). Une fois faite l'imputation des ventes du mois précédent, la valeur des ventes du mois courant est établie à partir de la valeur imputée pour le mois précédent, à l'aide de la méthode décrite relativement aux unités déclarantes choisies systématiquement.

Si une entreprise non répondante est incluse dans l'échantillon pour la première fois, on impute les ventes du mois précédent (dans le cas d'une entreprise choisie systématiquement) ou les ventes trois mois plus tôt (dans le cas d'une entreprise choisie aléatoirement) à partir des ventes déclarées au recensement le plus récent, si ces données existent. Lorsque l'entreprise non répondante n'a pas participé au recensement le plus récent, il s'agit alors d'une entreprise qui vient d'entrer dans l'échantillon, pour laquelle les données sur les ventes de deux mois ont généralement été fournies au moment où l'entreprise a été ajoutée à la base de sondage. Ces données sont alors désaisonnalisées et gonflées de façon à produire une valeur annuelle. L'imputation est alors effectuée comme si l'entreprise non répondante avait fourni la valeur de ses ventes lors d'un recensement.

4.2.2 Enquête sur le camionnage (Truck Inventory and Use Survey TIUS)

La TIUS est menée toutes les cinq années et vise à produire des données sur les caractéristiques physiques et l'exploitation des camions dans tout le pays. Ces caractéristiques comprennent le type de remorque (configuration du véhicule), les genres de produits transportés, le genre de carburant utilisé ainsi que le nombre de milles parcourus par véhicule au cours de l'année. L'univers de l'enquête est composé des immatriculations des camions dans les cinquante États et le district de Columbia. L'échantillon comprend environ 120,000 immatriculations de camions. Le taux de réponse des unités choisies pour participer à l'enquête est d'environ 75%.

Pour compenser la non-réponse au questionnaire, on augmente par pondération, séparément pour chaque classe de pondération, les réponses fournies par les répondants afin d'obtenir l'équivalent pour l'échantillon total. Les classes de pondération correspondent aux strates de l'échantillon qui sont composées de classifications combinées selon l'État et le type de caisse (5 catégories). La compensation de la non-réponse par pondération est basée sur le nombre de camions; dans chaque classe (strate) le poids initial de chaque répondant est multiplié par le ratio du nombre de camions dans la strate et de la somme des poids initiaux des répondants dans la strate.

De toutes les enquêtes économiques que nous avons étudiées, seule la TIUS utilise une procédure de correction de poids pour tenir compte de la non-réponse au questionnaire. Dans les autres enquêtes économiques, on dispose généralement d'autres sources de renseignements de base qui permettent de "créer" un enregistrement pour un non-répondant.

4.3 Recensement de l'agriculture

Le recensement de l'agriculture fournit des données sur les exploitations agricoles (cultures, élevages) et sur les activités connexes, à l'échelle nationale. C'est la source principale de statistiques agricoles et la seule source de données agricoles comparables au niveau des comtés, des États et du pays.

La compensation de la non-réponse dans le recensement de l'agriculture est complexe parce qu'on ne peut utiliser le SSEL aussi efficacement dans ce cas que dans les autres secteurs économiques. La liste d'envoi postal pour le recensement de l'agriculture est composée à partir de plusieurs sources qui se répètent. La base de sondage peut contenir par conséquent un certain nombre d'éléments en double et elle contient toujours certaines entités non agricoles. Aussi, la méthode de compensation de la non-réponse doit tout d'abord permettre de déterminer et d'estimer l'importance de la correction à apporter avant qu'on ne puisse l'effectuer.

4.2 Enquêtes économiques

Le Census Bureau effectue un grand nombre d'enquêtes économiques mensuelles, trimestrielles et annuelles, en plus des recensements économiques. Des enquêtes mensuelles ou annuelles sont menées pour la plupart des six secteurs commerciaux du recensement. Les méthodes utilisées pour la compensation de la non-réponse au questionnaire dans le Monthly Retail Trade Survey (enquête mensuelle sur le commerce au détail) et le Truck Inventory and Use Survey (enquête sur le camionnage) sont résumées plus loin. On ne décrit pas la méthode de compensation de la non-réponse au questionnaire utilisé pour l'Annual Survey of Manufacturers (enquête annuelle auprès des fabricants, ASM) parce que c'est virtuellement identique à celle qui est utilisée pour le Census of Manufacturers, dont la description a été donnée à la section 4.1.2. Les taux d'imputation pour l'ASM varient de 5 à 10%.

4.2.1 Enquête mensuelle sur le commerce de détail (Monthly Retail Trade Survey)

L'enquête mensuelle sur le commerce de détail comprend environ 30,000 unités déclarantes dont près de 3,000 sont choisies systématiquement et 27,000 aléatoirement. Les unités choisies systématiquement doivent participer à l'enquête chaque mois, alors que ce n'est le cas que pour un tiers des unités choisies aléatoirement. Il faut donc expédier un questionnaire par la poste à environ 12,000 unités déclarantes chaque mois. Dans le cas d'une entreprise à établissements multiples visée par l'enquête, c'est un sous-échantillon des établissements de cette entreprise qui est choisi pour être inclus dans l'enquête. Les chiffres mensuels de ventes au détail sont les seules données recueillies dans l'enquête. Le taux d'imputation pour ces ventes est d'environ 11%.

Si une entreprise à établissement unique choisie systématiquement ou un établissement sélectionné d'une entreprise à établissements multiples ne fait pas de déclaration un mois donné, on impute une valeur pour les ventes en multipliant les données du mois précédent par une facteur dit "rapport d'éléments identiques". Pour obtenir ce rapport de correction, il faut diviser la somme pondérée des ventes du mois courant par la somme pondérée des ventes du mois précédent pour tous les établissements de la même cellule de correction qui ont déclaré des ventes pour le mois courant et le mois précédent. Les cellules de correction sont généralement définies d'après les trois premiers chiffres (ou les quatre premiers chiffres dans quelques cas) du code SIC, selon le genre d'établissement (c.-à-d. si ce dernier appartient ou non à une grosse entreprise à établissements multiples), et selon la classe de l'échelle des ventes. Le poids utilisé pour chaque unité déclarante dans le calcul du facteur "rapport d'éléments identiques" est l'inverse de la probabilité de sélection de l'unité déclarante.

Si une entreprise à établissements multiples choisie systématiquement ne déclare pas de chiffres de ventes pour aucun de ses établissements, les valeurs des ventes sont imputées pour chaque établissement et pour l'entreprise comme dans le cas précédent: on applique le facteur "rapport d'éléments identiques", pour les cellules de correction correspondantes, aux données des ventes du mois précédent. Si une entreprise de ce genre déclare les ventes mensuelles courantes pour toute l'entreprise, les réponses imputées pour les établissements sont corrigées à l'aide d'un ratio de façon que le résultat soit compatible avec le total déclaré pour l'entreprise. Pour les entreprises choisies aléatoirement, l'imputation des données se fait comme dans le cas des unités déclarantes choisies systématiquement, sauf qu'une étape supplémentaire est nécessaire car les entreprises choisies aléatoirement produisent une déclaration tous les trois mois seulement. La première étape consiste à imputer le chiffre des ventes du mois précédent d'une entreprise non répondante d'après la réponse fournie trois mois plus tôt. Pour ce faire, on doit multiplier les ventes déclarées trois mois auparavant par le "rapport d'éléments

tionné plus haut. Le total relatif à l'entreprise est ensuite réparti entre les établissements proportionnellement au total obtenu à partir des données les plus récentes tirées d'une enquête annuelle ou mensuelle. S'il n'existe pas de données de ce genre, on procède à une répartition égale. S'il y a non-réponse pour seulement une partie des établissements d'une entreprise à établissements multiples, les données sur les établissements non répondants sont imputées selon les ratios qui existaient l'année précédente.

4.1.2 Fabrication, industries extractives

Dans ces deux recensements économiques, on obtient des renseignements généraux sur le nombre d'employés, les heures travaillées et les niveaux de production selon les codes SIC à quatre chiffres. Les taux d'imputation varient d'environ 10 à 15%. Les méthodes de compensation de la non-réponse au questionnaire varient selon le genre d'entreprise qui n'a pas répondu (c.-à-d., entreprise à établissement unique ou entreprise à établissements multiples) et selon qu'on dispose ou non d'un enregistrement pour l'année précédente. Ainsi, il y a quatre cas différents de non-réponse. Les méthodes de compensation de la non-réponse dans ces quatre cas sont:

- 1) Entreprise à établissement unique, les données de l'année antérieure sont tirées du Annual Survey of Manufactures.
- Dans ce cas, on obtient le chiffre de paye annuel à partir des déclarations d'impôt et on le compare au chiffre déclaré l'année précédente. Cette comparaison permet de déterminer le pourcentage de variation par rapport à la période antérieure. On applique ensuite ce pourcentage à tous les éléments d'information de l'enregistrement antérieur pour obtenir un enregistrement imputé (sauf pour l'emploi et la valeur des livraisons quand ces dernières valeurs peuvent être obtenues de l'IRS.)

- 2) Entreprise à établissement unique, aucune donnée relative à l'année antérieure.
- Dans ce cas, on utilise la paye comme "valeur de départ" pour calculer un ensemble de ratios entre les données du recensement, pour chaque entreprise codée à quatre chiffres dans la SIC. C'est-à-dire qu'on calcule des rapports de telle façon qu'on puisse imputer toutes les données à partir de ces rapports soit directement, soit indirectement, si on obtient le chiffre de paye. Les rapports particuliers sont obtenus à partir de données chronologiques déclarées par les répondants dans la même industrie. La valeur de la paye (valeur de départ) est alors tirée des dossiers fiscaux (IRS) et toutes les autres données sont imputées à partir des rapports obtenus.

- 3) Entreprise à établissements multiples, on dispose des données de l'année antérieure.
- On calcule tout d'abord, pour chaque entreprise codée à quatre chiffres, un facteur global de croissance entre la période antérieure et la période courante, à partir de sources externes pour chacune des données clés suivantes: paye, emploi, variation des stocks et variation des dépenses en immobilisations. Ces quatre facteurs de croissance sont appliqués aux données correspondantes de l'année précédente sur chaque établissement afin d'obtenir des réponses imputées pour la période en cours. Ces quatre données imputées sont alors utilisées comme "valeurs de départ" pour imputer d'autres données.
- 4) Entreprise à établissements multiples, aucune donnée de l'année précédente sur l'en-

Dans ce cas, on obtient les données de base sur la paye et l'emploi de chaque établissement à partir du SSEL dont on a parlé plus tôt à la section 4.1. Comme nous l'avons indiqué, le fichier SSEL contient des données sur l'emploi et la paye relativement à tous les établissements visés par la COS. Puis, en utilisant les données du SSEL comme base, on impute l'enregistrement de données sur chaque établissement à partir des rapports calculés entre les données du SSEL et les autres données du recensement. Cette procédure est analogue à celle utilisée pour le cas 2) ci-dessus.

sur les entreprises à établissements multiples. Les entreprises qui emploient au moins 50 personnes participent à cette enquête chaque année, tandis que les entreprises qui emploient moins de 50 personnes y participent tous les trois ans. On envoie à chaque entreprise visée une liste des établissements qu'elle a déclarés au cours de l'enquête la plus récente et on lui demande de la mettre à jour. L'entreprise doit aussi déclarer, pour chaque établissement, les effectifs au cours du premier trimestre de l'année précédente ainsi que le montant total de la paye versée l'année précédente. Pour les recensements économiques, on envoie à chaque établissement (par le biais de son siège social) qui figure dans le fichier SSEL, sauf les petites entreprises à établissement unique, un questionnaire de recensement qui a été conçu selon les caractéristiques des diverses classes de la classification américaine des activités économiques (code SIC).

Bien que les méthodes de compensation de la non-réponse au questionnaire pour les six secteurs commerciaux se ressemblent, il existe certaines différences importantes. Dans la description suivante des méthodes de compensation utilisées pour cinq des six recensements économiques, nous avons groupé les secteurs commerciaux pour lesquels on utilise essentiellement la même méthode:

- a) Commerce de détail, commerce de gros, services
- b) Fabrication, industries extractives

4.1.1 Commerce de détail, commerce de gros, services

Ces trois recensements économiques sont souvent considérés comme formant ensemble le recensement du commerce. Pour ces secteurs commerciaux, on recueille, relativement à l'année de recensement, des données sur les ventes et les recettes, l'emploi et la paye. Le taux d'imputation pour les ventes et les recettes varie de 10 à 15% dans le commerce de gros et de détail, et il est d'environ 20% dans les services.

Les réponses, pour tout établissement qui ne fournit pas les données requises dans le recensement, sont généralement imputées à partir des données fournies dans la déclaration d'impôt et stockées dans les fichiers du Internal Revenue Service (IRS). Pour les renseignements sur la paye, l'IRS dispose, à partir des déclarations d'impôt, de données pour quatre trimestres pour chaque numéro d'identification de l'employeur (numéro EI). Une entreprise peut avoir plusieurs numéros EI. Pour obtenir les données sur la paye d'une entreprise particulière, il faut additionner les chiffres de paye pour chacun des numéros EI utilisés par cette entreprise. On peut aussi obtenir, à partir des dossiers de l'IRS, le nombre d'employés par trimestre pour chaque numéro EI; ces chiffres peuvent aussi être groupés par entreprise. Pour les ventes et les recettes, l'IRS dispose de diverses déclarations d'impôt utilisées selon que l'entreprise non répondante est une entreprise individuelle, une société en nom collectif ou une société constituée.

L'imputation se complique du fait que l'unité de recensement est différente de l'entité fiscale de l'IRS. Pour le recensement du commerce, l'unité déclarante est l'établissement (c.-à-d. un local d'affaires). Pour l'IRS, l'unité déclarante est un numéro EI. Il se peut qu'un établissement ou plus produisent leurs déclarations d'impôt selon le même numéro EI. Si une entreprise non répondante n'a qu'un local d'affaires (c.-à-d. s'il s'agit d'une entreprise à établissement unique), il n'y aura qu'un seul numéro EI et l'imputation se fait sans problème particulier. Cependant, dans le cas d'une entreprise non répondante à établissements multiples, l'imputation est plus complexe puisque, en général, nous ne disposons pas de données de l'IRS sur chaque établissement. Dans un tel cas, on doit tout d'abord déterminer la structure de l'entreprise en se reportant au SSEL afin d'obtenir une liste de tous les établissements de cette entreprise et tous les numéros EI que cette dernière utilise. On obtient le total de chaque élément d'information sur une entreprise en additionnant les chiffres déclarés pour tous les numéros EI utilisés par cette entreprise, comme nous l'avons men-

simplifier le problème des données manquantes dans la SIPP; et 5) l'évaluation des méthodes de compensation de la non-réponse longitudinale établies pour le premier fichier longitudinal de recherche de la SIPP.

4. MÉTHODES DE COMPENSATION DE LA NON-RÉPONSE POUR LES RECENSEMENTS ET ENQUÊTES ÉCONOMIQUES

Le Bureau of the Census effectue six recensements économiques toutes les cinq années; les plus récents sont ceux qui portent sur l'année 1982. Ces six recensements économiques sont désignés par l'appellation des secteurs commerciaux suivants:

- 1) Commerce de détail
- 2) Commerce de gros
- 3) Services
- 4) Fabrication
- 5) Industries extractives
- 6) Construction

En plus des recensements économiques, le Census Bureau effectue un recensement des administrations publiques (Census of Government) et un recensement de l'agriculture qui sont menés les mêmes années que les recensements économiques afin de faciliter le traitement et de faire le couplage des données. Le Census Bureau effectue aussi un certain nombre d'enquêtes mensuelles, trimestrielles et annuelles dans presque tous ces secteurs économiques. Comme dans le cas des enquêtes démographiques, il y a un certain taux de non-réponse au questionnaire des recensements et enquêtes économiques. Dans la plupart des cas, les données omises sont imputées d'après a) des réponses déjà fournies par le non-répondant, b) des données tirées de dossiers administratifs et c) des liens établis entre diverses données élémentaires. Plutôt que de déclarer le pourcentage d'unités qui n'ont pas répondu, on indique habituellement le niveau de non-réponse dans un recensement ou une enquête économique sous forme de pourcentage d'un ou de plusieurs totaux imputés pour des questions. On désigne ces pourcentages les taux d'imputation.

Des explications sur les méthodes de compensation de la non-réponse au questionnaire utilisées pour cinq des six recensements économiques sont données à la section 4.1. Dans la section 4.2, on traite des méthodes de compensation de la non-réponse dans trois enquêtes économiques et, à la section 4.3, des méthodes équivalentes utilisées dans le recensement de l'agriculture. Les travaux de recherche et d'évaluation portant sur les méthodes de compensation de la non-réponse dans les recensements et enquêtes économiques sont examinés à la section 4.4.

Pour des explications plus détaillées sur les méthodes de compensation de la non-réponse dans ces recensements et dans plusieurs enquêtes connexes, voir Bailey, Chapman et Kasprzyk (1985).

4.1 Les recensements économiques

La base de sondage utilisée pour les recensements économiques est la Standard Statistical Establishment List (SSEL), un fichier informatique tenu à jour par le Census Bureau. La SSEL est un répertoire de tous les établissements employeurs déclarés par les entreprises employées à établissements multiples dans le cadre de la Company Organization Survey (COS) du Census Bureau, ainsi que de toutes les entreprises à établissement unique qui ont produit une déclaration d'impôt auprès de l'IRS. La COS est une enquête annuelle portant

À mesure que l'enquête progresse, des méthodes plus perfectionnées de compensation de la non-réponse longitudinale seront élaborées en fonction des données fournies par les répondants qui fournissent une partie des renseignements demandés (c.-à-d. pour les personnes échantillonnées qui répondent à certaines, et non à toutes, les interviews des différents cycles auxquels elles sont supposées participer). Le traitement des cas de réponse partielle n'est pas évident. Les données manquantes dues au fait que des personnes n'ont pas répondu à une interview ou plus peuvent être considérées soit comme une non-réponse par une personne, cas généralement traité par des corrections de poids, soit comme une non-réponse à une question, cas généralement traité par imputation. Par exemple, on pourrait considérer le cas d'une personne dont la présentation des réponses serait de type (R, NR, R) comme un cas de non-réponse à une question car l'interview manquante est encadrée par des interviews achevées; mais on pourrait considérer le cas d'une personne dont la présentation des réponses serait de type (NR, R, NR) comme un cas de non-réponse au questionnaire, et le traiter de la même façon qu'un cas de présentation des réponses de type (NR, NR, NR). Nous devons cependant admettre que même dans le cas où la présentation des réponses est du type (R, NR, R) selon la personne, il est encore possible de trouver quatre types de présentations des réponses selon la question. On peut donc considérer de nombreuses possibilités au moment de l'élaboration des méthodes de compensation de la non-réponse pour la base de données longitudinale de la SIPP. Cette question est traitée par Kalton (1986) et par Kalton, Lepkowski, et Lin (1985).

3.2 Recherches relatives à la SIPP

Des travaux entrepris récemment dans deux domaines pourraient aider à la prise de décisions futures concernant la compensation de la non-réponse. Tout d'abord, à compter de la quatrième interview, le questionnaire de la SIPP contient une section sur le "cycle manquant" ("Missing Wave"). Cette section comprend un petit nombre de questions sur l'activité, les sources de revenu et la possession ou non-posssession d'actifs, qui doivent être posées aux répondants dans le cycle courant qui n'ont pas répondu au questionnaire dans le cycle précédent. Les répondants qui n'ont pas répondu à deux interviews consécutives ou plus ne sont pas autorisés à remplir la section sur le "cycle manquant". On peut donc, en accordant plus d'importance à la collecte des données, même si cela comporte une légère augmentation du fardeau de déclaration, réduire le problème de la non-réponse due à une personne à un problème de non-réponse à une question. Il faudra cependant évaluer la qualité des données chronologiques avant d'utiliser ces données.

Le second domaine dans lequel des travaux sont en cours est celui des techniques générales de traitement de la non-réponse d'une personne au cours d'un cycle de la SIPP. Graham Kalton et ses collègues du Survey Research Centre doivent 1) comparer les techniques d'imputation et de pondération longitudinales pour le traitement de la non-réponse d'une personne au cours d'un cycle, 2) évaluer les modèles d'imputation et de pondération en fonction de l'analyse du changement entre les cycles et du groupement entre les cycles et 3) définir des critères provisoires pour le choix d'une méthode de traitement de la non-réponse d'une personne au cours d'un cycle. On trouve une analyse de ces questions et d'autres qui seront traitées plus loin dans Kalton (1986), et Kalton et Miller (1986).

Enfin, des recherches sont prévues pour plusieurs autres sujets, notamment 1) l'attribution d'une valeur numérique au choix des variables utilisées pour déterminer les classes de pondération; 2) l'évaluation de la robustesse des estimations de l'enquête relatives à la population et à certains sous-groupes selon diverses méthodes de compensation de la non-réponse et diverses techniques d'agrégation des cellules des classes de pondération; 3) l'étude de la possibilité d'effectuer des compensations distinctes de la non-réponse selon le genre de non-interview; 4) l'étude de l'effet de la suppression de données déclarées dans une enquête afin de

Tableau 6

Présentations des interviews des personnes incluses dans l'échantillon initial pour les cinq premières interviews du panel de la SIPP de 1984

Présentation des réponses		Pourcentage
Réponse à toutes les interviews (5 interviews)		
Présentation: XXXXX		
Cas apparents de perte d'effectifs		
Présentations: XXXXO		
3.8		
3.1		
3.2		
3.7		
Les première et cinquième interviews ont été complétées, mais au moins une des interviews intermédiaires manque		
Présentations: XXXOX		
1.6		
0.6		
1.2		
0.1		
0.1		
0.3		
0.2		
La cinquième interview manque et au moins une des deuxième à quatrième interviews manque aussi		
Présentations: XXXOX, XXXOX, XXXOX, XXXOX		
0.7		
A quitté l'univers (décédé, placé dans un établissement, habite dans des casernements des Forces armées, a déménagé outre-mer)		
2.3		
Total		
100.0		
(25,128)		

Pour le premier facteur de compensation, on peut utiliser seulement les variables des ménages, qui ont été observées à la première interview. Les facteurs de compensation sont calculés séparément à l'intérieur des cellules définies par les variables suivantes:

- a. Région de recensement
- b. Domicile (région métropolitaine, région non métropolitaine)
- c. Race de la personne-repère
- d. Mode d'occupation (propriétaire, locataire)
- e. Taille du ménage

Le deuxième ensemble de facteurs de compensation est appliqué au niveau des personnes. Les facteurs sont calculés à l'intérieur des cellules définies par les caractéristiques suivantes:

- a. Revenu mensuel du ménage
- b. Situation du ménage de la personne sur le plan de la participation au programme
- c. Situation vis-à-vis de l'activité
- d. Race
- e. Nombre d'années d'études terminées
- f. Type d'actifs détenus par le ménage de la personne

Les cellules sont regroupées chaque fois qu'elles ne contiennent pas trente personnes échantillonnées ou lorsque le facteur de compensation de la non-réponse est supérieur à 2.

6. Mode d'occupation : a) propriétaire de son domicile et b) locataire.
7. Logements sociaux ou allocations de logement – les locataires sont divisés en deux catégories a) ceux qui habitent des logements sociaux ou auxquels le gouvernement verse des allocations de logement et b) ceux qui n'habitent pas des logements sociaux ou auxquels le gouvernement ne verse pas d'allocation de logement.
8. Taille du ménage: 1, 2, 3, 4 personnes ou plus.

Les variables utilisées pour la compensation de la non-réponse des ménages à la deuxième interview de la SIPP et aux interviews subséquentes diffèrent des variables utilisées pour le premier cycle parce que, après la première interview, on dispose de données additionnelles qu'on applique au traitement de la non-réponse dans les interviews suivantes. Cinquante-trois classes de pondération ont été créées à partir de ces variables et c'est le mode d'occupation qui a été la principale variable de fractionnement de l'échantillon. [Pour une description de ces classes de pondération, voir U.S. Department of Commerce, Bureau of the Census (1984c).] Bien qu'on ait défini une méthode d'agrégation des cellules selon laquelle les cellules qui possèdent des caractéristiques semblables liées à la pauvreté sont fusionnées, il y a peu d'agrégation car les facteurs de compensation de la non-réponse sont calculés pour trois groupes de renouvellement (le cycle informatif de la SIPP) plutôt que pour un groupe de renouvellement, comme dans le cas de la première interview.

Il existe une technique de compensation de la non-réponse à l'intérieur des ménages pour le deuxième cycle et ceux qui suivent. Pour appliquer cette technique, qui est fondée sur la méthode "hot deck", il faut reproduire l'enregistrement complet d'un répondant dont on suppose que les caractéristiques sont semblables à celles du non-répondant.

3.1.2 Compensations de la non-réponse longitudinale

Comme les personnes déclarées habitant à l'adresse incluse dans l'échantillon au moment de la première interview constituent l'échantillon de la SIPP pour les cycles qui suivent le premier, c'est en fonction des individus ou des personnes qu'il est le plus utile et logique de décrire la nature du problème de la non-réponse longitudinale dans la SIPP. L'enregistrement des microdonnées de chaque personne est un enregistrement élargi qui contient des variables exprimant souvent la même mesure à un moment différent. Ainsi, dans une enquête par panel de n cycles, il existe 2ⁿ présentations possibles des non-interviews pour une personne échantillonnée. On trouve au tableau 6, les présentations des non-interviews des personnes qui faisaient partie de l'échantillon initial pour les cinq premières interviews (cycles) du panel de 1984, ces renseignements sont adaptés de Kalton, McMillen et Kasprzyk (1986). Le premier fichier longitudinal de microdonnées de la SIPP contiendra douze mois (trois interviews) de données provenant du panel de la SIPP de 1984, et la personne est la principale unité d'analyse. L'échantillon des cas à pondérer pour ce fichier comprendra seulement les personnes pour lesquelles trois interviews ont été effectuées. Les personnes échantillonnées pour lesquelles seulement une ou deux interviews ont été effectuées seront traitées comme des non-répondants. Les données qu'elles auront fournies aideront à définir les classes de compensation de la non-réponse.

Comme le premier fichier longitudinal de microdonnées concerne uniquement les personnes qui ont répondu aux trois interviews, la question de la correction pour la non-réponse est, à toutes fins pratiques, identique à celle pour le fichier transversal. Cependant, deux facteurs de compensation de la non-réponse sont appliqués aux poids de l'échantillon initial. [Voir Kobilarcik et Singh (1986).] Le premier facteur de compensation tient compte des ménages classés comme des cas de non-interview au premier cycle. Le second facteur tient compte des personnes qui n'ont pas répondu aux trois interviews.

3.1.1 Compensations de la non-réponse transversale au questionnaire

Pour la SIPP, la compensation de la non-réponse transversale au questionnaire est semblable à la façon dont les corrections pour la non-interview sont effectuées dans les autres enquêtes périodiques du Census Bureau. Les variables utilisées pour définir les cellules de correction pour les cas de non-interview des ménages dans le premier cycle d'interview de la SIPP sont les suivantes [voir U.S. Department of Commerce, Bureau of the Census (1983 et 1984b)]:

1. Région de recensement – Nord-est, Midwest, Sud, Ouest.
2. Résidence – région métropolitaine type (Standard Metropolitan Statistical Area, SMSA), hors d'une SMSA.
3. Localité/hors localité – pour les unités qui ne sont pas dans une SMSA.
- Principale ville / autres agglomérations – pour les unités dans une SMSA.
4. Race de la personne-repère – Noir, Non-Noir.
5. Mode d'occupation – propriétaire de son domicile, locataire.
6. Taille du ménage – 1, 2, 3, 4 personnes ou plus.
7. Groupe de renouvellement – 1, 2, 3, 4.

Chaque classe de pondération doit satisfaire à deux conditions: 1) elle doit comprendre au moins 30 unités non pondérées et 2) le facteur de compensation de la non-réponse, dans le cas d'une classe de pondération, doit être inférieur ou égal à 2.0. Pour un groupe de renouvellement donné, la procédure d'agrégation visant à satisfaire à ces deux conditions est appliquée indépendamment pour chacune des quatre combinaisons de mode d'occupation selon la race. (Pour le premier cycle, il n'y avait pas de facteur de compensation de la non-réponse à l'intérieur des ménages.)

Dans les cycles ultérieurs de la SIPP, le facteur de compensation de la non-réponse tient compte des non-interviews des unités qui ont déménagé et qui ne peuvent être retracées ou qui ont déménagé à plus de 100 milles d'une U.P.E. de la SIPP et qu'on ne peut joindre par téléphone, ainsi que des unités qui refusent de répondre, etc. Les rajustements sont effectués pour chaque mois de la période de référence ainsi que pour le mois d'interview, afin de tenir compte de l'augmentation du nombre de non-interviews qui est due au fractionnement des ménages dans l'échantillon. La procédure est semblable à celle qui est utilisée pour déterminer le facteur de compensation de la non-réponse des ménages au 1^{er} cycle; cependant, les variables utilisées pour définir les classes de pondération diffèrent. Ces variables sont:

1. Race (Blanc, Non-Blanc) et origine hispanique (hispanique, non hispanique) de la personne-repère: a) la personne-repère est blanche et non hispanique, et b) autres.
2. Genre de ménages – trois catégories: a) femme responsable du ménage, sans conjoint, avec ses propres enfants de moins de 16 ans, b) membre responsable du ménage âgé de 65 ans ou plus, et c) autres.
3. Niveau de scolarité de la personne-repère: a) moins de 8 années, b) de 8 à 11 années, c) de 12 à 15 années, et d) 16 années ou plus.

4. Genre de revenu reçu (d'après la plus récente interview menée auprès des membres du ménage) – deux catégories: a) ménages ayant reçu un revenu d'au moins une des sources suivantes – Supplemental Security Income; Black Lung Payments; Aid to Families with Dependent Children; General Assistance; Indian Assistance; Cuban Assistance ou Refugee Assistance; paiements pour la garde d'un enfant en foyer nourricier; Women's, Infants', and Children's Nutrition Program; Food Stamps; et Medicaid; et b) autres.
5. Éléments d'actifs – deux catégories: a) ménages dans lesquels au moins un membre détient des éléments d'actifs autres qu'un compte d'épargne ou qu'un compte de chè-

ques qui porte intérêt, et b) tous les autres.

Tableau 5
Taux de non-interview cumulatifs des ménages pour les
panels de la SIPP de 1984

Cycle	Perte d'effectifs de l'échantillon
1	4.9%
2	9.4%
3	12.3%
4	15.4%
5	17.4%
6	19.4%
7	21.0%
8	22.0%
9	22.3%

si elles déménagent au cours des deux années et demie suivantes. Pour des raisons d'ordre opérationnel et budgétaire, les interviews sur place sont effectuées aux nouvelles adresses seulement si ces dernières se trouvent à moins de 100 milles d'une unité primaire d'échantillonnage de la SIPP. Plus de 96% de la population des Etats-Unis habite dans les régions géographiques établies de cette façon. Pour les personnes qui ont déménagé à l'extérieur de cette limite de 100 milles, on essaie d'effectuer une interview téléphonique.

Après la première interview, l'échantillon de la SIPP est un échantillon de personnes, qui comprend toutes les personnes qui habitaient l'unité d'échantillonnage au moment de la première interview. Les personnes âgées de 15 ans ou plus qui, par la suite, partagent des pièces d'habitation avec les personnes qui faisaient partie de l'échantillon au début sont aussi interviewées afin qu'on puisse définir le contexte économique global des personnes qui faisaient partie de l'échantillon initial.

Pour plus de renseignements sur le plan de sondage, le contenu et les opérations relatives à la SIPP, voir l'ouvrage de Nelson, McMillen et Kasprzyk (1985).

3.1 Compensations de la non-réponse dans la SIPP

Les données recueillies dans le cadre de la SIPP peuvent être étudiées de deux points de vue; elles peuvent servir à des études transversales ou à des études longitudinales. Dans le premier cas, chaque interview effectuée pour la SIPP est traitée comme une enquête transversale distincte qui fournit des estimations ponctuelles. Pour des exemples de ces estimations, voir U.S. Department of Commerce, Bureau of the Census (1984a). Du point de vue longitudinal, les données sont recueillies à plusieurs moments donnés et chaque enregistré-moment de l'enquête est considéré non comme un ensemble d'observations sans liens entre elles, mais comme un ensemble de variables entre lesquelles il existe une dépendance logique à deux moments ou plus. Comme le traitement informatique et l'estimation statistique sont effectués en fonction des analyses longitudinales, ces opérations sont basées sur l'utilisation de données recueillies dans deux interviews ou plus.

Puisqu'on peut considérer la SIPP dans une perspective d'analyses tant longitudinales que transversales, les fichiers de microdonnées à grande diffusion comprennent des fichiers de données transversales produits cycle par cycle ainsi que des fichiers longitudinaux. Cela suppose donc deux systèmes distincts de traitement de la non-réponse à l'enquête.

Thomsen et Sirling (1983). Il se peut que ces méthodes soient applicables aux enquêtes périodiques pour lesquelles beaucoup de rappels sont effectués.

Des recherches sont en cours sur la construction de modèles qui pourraient servir, pour plusieurs enquêtes démographiques, à l'estimation des probabilités de réponse des unités pour lesquelles les "variables indépendantes" ont des valeurs semblables. On étudie la faisabilité et les mérites de calculer des facteurs de compensation de la non-réponse ainsi que de construire des classes de pondération basées sur de tels modèles (désignées parfois stratification de la propension à répondre). [Voir Rosenbaum et Rubin (1983) et Little et Samuël (1983).] De plus, des travaux se poursuivent en vue d'élaborer des méthodes plus objectives de pondération de l'échantillon, et ce afin de contrôler les erreurs liées à la non-réponse.

3. L'ENQUÊTE SUR LE REVENU (SURVEY OF INCOME AND PROGRAM PARTICIPATION)

L'enquête sur le revenu aux Etats-Unis ou Survey of Income and Program Participation (SIPP) est un nouveau programme d'enquête national permanent mené auprès des ménages par le U.S. Bureau of the Census. L'objet de la SIPP est d'améliorer la mesure de l'information sur la situation économique des ménages et des particuliers aux Etats-Unis. C'est l'aboutissement d'un vaste programme d'élaboration, l'Income Survey Development Program (ISDP), dans le cadre duquel on a examiné les définitions, les méthodes, les questionnaires, les périodes visées par l'enquête et autres éléments du genre. [Pour une description du ISDP, voir Ycas et Lininger (1981).] On s'attend que les données obtenues dans le cadre de la SIPP serviront à l'analyse du système de transfert fédéral, à l'estimation du coût des programmes si des modifications sont apportées aux critères d'admissibilité, à l'évaluation des effets des changements apportés aux programmes sur certains sous-groupes de la population ainsi qu'à l'examen des modifications apportées au régime fiscal.

La SIPP a été lancée en octobre 1983 comme programme d'enquête permanent. Elle comportait un panel d'environ 21,000 ménages de logements occupés admissibles à participer à une interview dans 174 unités primaires d'échantillonnage (U.P.E.) choisies pour représenter les membres de la population des Etats-Unis qui n'habitent pas dans un établissement. À compter de 1985, un nouveau panel est inclus dans l'enquête, en février de chaque année; le panel de 1985 comprenait 14,500 ménages admissibles à une interview.

Chaque ménage est interviewé une fois tous les quatre mois, pendant environ 2 1/2 ans, dans le but de recueillir suffisamment de données pour pouvoir effectuer une analyse longitudinale tout en fixant une période relativement courte pour la déclaration du revenu mensuel. La période de référence des principales questions de l'enquête correspond aux 4 mois qui précèdent l'interview. Ce plan de sondage permet d'effectuer huit interviews par ménage et aussi de produire des estimations transversales à partir des réponses de plus d'un panel.

Pour faciliter le travail sur le terrain et le traitement, chaque panel de l'échantillon est divisé en quatre sous-échantillons comprenant approximativement le même nombre de ménages; ce sont des groupes de renouvellement. Un groupe de renouvellement est interviewé chaque mois. Ainsi, un cycle d'interviews, faites à l'aide du même questionnaire, prend quatre mois consécutifs. Les taux de non-interview cumulatifs des ménages sont donnés au tableau 5 pour les panels de la SIPP de 1984.

Au moment de la visite de l'interviewer, on demande à chaque personne présente, âgée de 15 ans ou plus, de fournir des renseignements à propos d'elle-même; on demande à un membre du ménage de fournir les renseignements pour les absents. Une caractéristique importante du plan de sondage de la SIPP, c'est que toutes les personnes qui font partie d'un ménage échantillonné au moment de la première interview demeurent dans l'échantillon même

Le choix des classes de pondération selon cette méthode est limité par le fait que des mesures pour les variables de la classe de pondération (covariables) doivent être produites (soit avant, soit au cours de l'enquête) pour les répondants ainsi que pour les non-répondants. Cela limite essentiellement les caractéristiques utilisées pour définir les classes par rapport à celles concernant la géographie, la race, le domicile dans une ville ou non, les caractéristiques du logement et les niveaux de plan de sondage. Les moyens qu'offre cette méthode pour réduire les biais dépendent, en partie, de la mesure dans laquelle les classes de pondération pour la non-réponse, dans la NCS, confirment les trois hypothèses données plus haut. Aucun résultat définitif relatif à ce sujet n'est connu actuellement, mais des travaux de recherche sont en cours et il semble que d'autres études empiriques sont justifiées.

2.2 Autres méthodes possibles de compensation

Il existe un certain nombre d'autres méthodes qui pourraient être utilisées, à la place de la pondération, pour corriger la non-réponse (voir par exemple Little, 1986, chapitre 5). Cependant, il n'existe pas de résultats définitifs qui montrent qu'une de ces autres méthodes offre des avantages appréciables. Les sections 2.2.1 et 2.2.2 décrivent sommairement deux autres méthodes qui font présentement l'objet d'analyses en vue de leur application dans des enquêtes démographiques.

2.2.1 Estimations distinctes pour des types de non-réponse différents

Dans les enquêtes démographiques, on peut diviser les non-répondants en quatre catégories: refus (REF), absent (ABS), autre logement occupé (ALO), ou logement où il a été impossible d'obtenir une réponse à cause de circonstances spéciales. Pour décrire ces cas de non-réponse, on utilise l'expression non-interview de type A. Le groupe ABS peut être divisé en ménages ou personnes dont l'absence prolongée de leur domicile empêche toute interview au cours de la période prévue pour les interviews (ABS_L) (absence de longue durée), et le groupe des personnes dont le retour est prévu au cours de la période d'enquête (ABS_C) (absence de courte durée).

Les auteurs ne connaissent pas de données qui démontrent que les quatre groupes de non-réponse sont généralement semblables. En fait, les résultats de la Current Population Survey du Census Bureau et ceux de l'Enquête sur la population active, effectuée au Canada, semblent indiquer que les ménages ABS_C sont vraisemblablement plus petits, plus jeunes et qu'ils comprennent une plus grande proportion de personnes occupées que les autres groupes. Le groupe ABS_L est habituellement plus âgé et son taux d'emploi est relativement faible. Le groupe interviewé peut mieux correspondre au groupe REF et ALO. [Voir Palmer et Jones (1967) et Paul et Lawes (1982).] Il est concevable qu'un traitement distinct des quatre groupes de non-réponse puisse mieux compenser les cas de non-réponse que ce que fait la méthode courante. Cette possibilité fait l'objet d'une étude par un groupe de recherche sur les méthodes de compensation de la non-réponse pour la NCS.

2.2.2 Pondération avec des probabilités de réponse

Plusieurs techniques de pondération où l'on applique le concept de probabilité de réponse ont été proposées. La plupart de ces techniques sont basées sur des définitions établies par Politz et Simmons (1949) qui regroupent les répondants dans l'échantillon selon les estimations de leurs probabilités de répondre. Le facteur utilisé pour gonfler les données-échantillons afin d'obtenir les groupes de pondération est l'inverse de la probabilité de réponse estimée. La méthode de Politz-Simmons comporte des limites importantes, comme le fait qu'elle ne peut être appliquée au traitement des refus. Cependant, un certain nombre de développements et d'applications de la procédure ont été faits récemment, notamment par Anderson (1978) et

Tableau 4
Taux de non-interview dans la NCS - 1984

Average 1984	janv.	fév.	mars	avril	mai	juin
-----------------	-------	------	------	-------	-----	------

Non-interview des ménages						
Nombre total de chefs de ménage interviewés	11,769	11,916	11,925	11,743	11,809	11,918
Total	430	446	540	481	446	388
Taux	3.5	3.6	4.3	3.9	3.6	3.2
Personne à domicile	0.9	0.8	1.1	0.9	0.9	0.7
Absent temporairement	0.6	0.6	0.6	0.8	0.6	0.4
Refus	1.9	2.1	2.6	2.2	2.2	2.0
Autre	0.1	0.2	0.2	0.1	0.1	0.1
Non-interviews à l'intérieur des ménages						
Total	685	655	751	701	806	804
Taux	2.5	2.6	3.0	2.8	3.0	2.9

Non-interview des ménages						
Nombre total de chefs de ménage interviewés	9,869	9,446	9,895	9,350	9,692	9,410
Total	411	409	337	406	387	346
Taux	4.0	4.2	3.3	4.2	3.8	3.5
Personne à domicile	0.9	0.9	0.6	1.0	1.2	1.0
Absent temporairement	1.0	1.0	0.6	0.6	0.4	0.4
Refus	2.1	2.3	2.0	2.4	2.1	2.1
Autre	0.1	0.1	0.1	0.3	0.3	0.1
Non-interviews à l'intérieur des ménages						
Total	709	678	666	728	735	803
Taux	3.1	3.1	2.9	3.4	3.3	3.7

Les hypothèses suivantes sont faites implicitement pour la formation des classes de pondération pour la non-réponse dans la NCS ainsi que pour celles d'autres enquêtes démographiques:

1. Il existe une corrélation "importante" entre les principales variables d'enquête et les covariables utilisées pour définir les classes voisines pour la non-interview.
2. Dans chaque classe de pondération pour la non-réponse des ménages, $E(\bar{y}_{Rj}) = E(\bar{y}_{Rj})$, où \bar{y}_{Rj} et \bar{y}_{Rj} sont les moyennes pour les répondants et les non-répondants dans l'échantillon, respectivement, dans la j^{e} classe de pondération.
3. Les moyennes des classes de pondération diffèrent, c'est-à-dire, $E(\bar{y}_{Rj}) \neq E(\bar{y}_{Rj'})$, $j \neq j'$.

(Des hypothèses semblables sont aussi faites implicitement pour les classes de compensation de la non-réponse à l'intérieur des ménages.)

Tableau 1
Cellules de correction pour la non-interview dans les cas de non-réponse à l'intérieur des ménages

Lien avec les membres du ménage		Personnes selon l'âge, selon la race du chef					
		Noir			Non-noir		
Chef de ménage	Épouse du chef	Toutes les autres personnes	12-34	25-44	45-64	65 +	12-24
			25-44	45-64	65 +	25-44	45-64

Tableau 2
Cellules de correction pour la non-interview des ménages, pour la NCS, dans le cas des Standard Metropolitan Statistical Area (SMSA)

Race	Principale ville de la SMSA	Reste de la SMSA	
		Région urbaine	Région rurale

Tableau 3
Cellules de correction pour la non-interview des ménages, pour la NCS, dans le cas des régions autres qu'une SMSA

Race	Région urbaine	Région rurale	
		Non agricole	Agricole

Pour illustrer l'estimateur d'un total de la NCS, posons une probabilité de sélection, $\pi_i = 1, 2, \dots, N$, appliquée à chacune des N unités dans la population. On suppose que parmi les n unités d'échantillonnage, n_R sont des unités répondantes. L'estimateur de la NCS pour l'ensemble de la population, peut être exprimé par la formule suivante, une fois la composition de la non-réponse au questionnaire effectuée:

$$Y_{NCS} = \sum_{M} \sum_{P}^{j=1} (z_j u_k) - \sum_{n_{Rjk}}^{\ell=1} \frac{\pi_{jkl}}{y_{jkl}}.$$

où pour les unités visées dans la k^e classe de pondération à l'intérieur des ménages et la j^e classe de pondération des ménages,

- y_{jkl} = valeur du répondant ℓ dans l'échantillon
- n_{Rjk} = nombre de répondants dans l'échantillon
- n_k = nombre de cas dans l'échantillon
- z_j = le taux de réponse des ménages estimé
- u_k = le taux de réponse à l'intérieur des ménages estimé,
- π_{jkl} = probabilité de sélection du répondant ℓ dans l'échantillon
- P = nombre total de classes de pondération pour la non-réponse à l'intérieur des ménages
- M = nombre total de classes de pondération pour la non-réponse des ménages.

En dépit des différences dans le contenu et la nature des enquêtes démographiques du Census Bureau ainsi que de l'importance de la non-réponse au questionnaire, la correction de poids à l'intérieur des classes (Oh et Scheuren 1983), aussi appelée compensation par cellule, est la principale technique utilisée pour compenser la non-réponse au questionnaire. On n'applique pas la même méthode d'établissement des classes de compensation pour la pondération. Dans certaines enquêtes, les données auxiliaires connues et utilisées pour définir les classes de pondération se limitent à des renseignements de base de nature géographique et l'information tirée du plan de sondage, tandis que pour d'autres enquêtes on peut disposer d'un volume considérable de données démographiques et économiques.

Les facteurs de compensation de la non-réponse établis pour les enquêtes démographiques du Census Bureau sont habituellement l'inverse du taux de réponse pondéré ou non de l'enquête. Dans un petit nombre d'enquêtes, ce facteur est légèrement modifié de manière qu'on puisse tenir compte des renseignements fournis dans les enquêtes de suivi menées auprès de sous-échantillons de personnes qui n'avaient pas répondu à l'enquête. Puisque la méthode généralement utilisée par le Census Bureau dans les cas de non-réponse à une enquête est essentiellement la même pour toutes ses principales enquêtes démographiques, nous donnons, à la section 2.1, une description générale de la méthode de compensation de la non-réponse utilisée pour la National Crime Survey (NCS), à titre d'exemple d'une application "typique" de la pondération au Census Bureau. La section 2.2 présente une analyse d'autres méthodes de compensation possibles ainsi que les travaux de recherche en cours sur la non-réponse aux enquêtes démographiques.

2.1 L'enquête nationale sur la criminalité (National Crime Survey)

L'échantillon de la NCS est un échantillon aléatoire composé d'environ 72,000 ménages à l'échelle nationale. Cet échantillon est divisé en six panels et chaque panel est interviewé un mois donné et à nouveau à des intervalles de six mois sur une période de trois ans. L'enquête vise à mesurer la criminalité ainsi que le nombre de fois où les membres d'un ménage âgés de 12 ans ou plus ont été victimes d'agression (y compris le viol), de cambriolage, de vol simple, de vol d'auto et de vol. [Pour une description détaillée de la NCS, voir U.S. Department of Commerce, Bureau of the Census (1977).]

Pour produire les estimations trimestrielles de la NCS, il faut tout d'abord gonfler les données-échantillons en les multipliant par l'inverse des probabilités de sélection correspondantes. Les cas où il n'y a pas eu de contact et les refus représentent entre trois et quatre pour cent des logements occupés inclus chaque mois dans l'enquête. Des corrections sont apportées pour tenir compte de ces unités: on applique des facteurs de compensation aux données pondérées sur les répondants dans les classes de pondération. Ces classes sont définies de telle sorte que les répondants et les non-répondants dans chaque classe ont des caractéristiques semblables. Afin de réduire l'effet de la compensation de la non-réponse sur la variance des estimations de l'enquête, il faut généralement combiner certaines des plus petites classes de pondération avec d'autres classes avant de pouvoir effectuer une compensation finale de la non-réponse. Il faut également combiner les classes lorsque les facteurs d'ajustement de la non-réponse de la NCS en est une où les répondants remplissent eux-mêmes le questionnaire, on se préoccupe de l'importance de la non-réponse à l'intérieur des ménages. Aussi, on a établi un ensemble distinct de cellules de pondération pour compenser la non-réponse à l'intérieur des ménages. Ces cellules ou classes de pondération, ainsi que les tableaux 1 à 3. Les taux de non-réponse des ménages, sont décrites dans les tableaux pour la NCS de 1984 figurent au tableau 4.

Méthodes de compensation de la non-réponse au U.S. Bureau of the Census DAVID W. CHAPMAN, LEROY BAILEY, et DANIEL KASPRZYK¹

RÉSUMÉ

Presque tous les recensements et enquêtes comportent deux types de non-réponses: la non-réponse au questionnaire (non-réponse totale) et la non-réponse à une question (non-réponse partielle). Plusieurs méthodes de compensation de la non-réponse ont été élaborées pour tenter de réduire la distorsion due à la non-réponse. Cet article résume les méthodes de compensation de la non-réponse utilisées au U.S. Census Bureau, et traite particulièrement du problème de la non-réponse au questionnaire. On examine aussi sommairement les travaux de recherche actuels et futurs dans ce domaine.

MOTS CLÉS: Compensation de la non-réponse; imputation; données manquantes; pondération.

1. INTRODUCTION

Le Bureau of the Census a admis depuis longtemps l'importance que peuvent avoir les erreurs de mesure imputables à la non-réponse dans les enquêtes et a toujours prévu des méthodes de compensation de la non-réponse dans les méthodes d'estimation utilisées pour les nombreux et divers recensements et enquêtes qu'il effectue. Cet article a pour objet de donner un aperçu des méthodes de compensation de la non-réponse utilisées par le Census Bureau, principalement dans le cas de la non-réponse au questionnaire. Par non-réponse au questionnaire nous entendons les cas où peu ou aucun renseignement n'a été obtenu, relativement aux principales variables d'enquête, pour l'unité d'échantillonnage visée. Cet exposé comprend 1) une analyse de la méthode de pondération générale appliquée dans les enquêtes démographiques, 2) un examen de quelques-uns des problèmes liés à la non-réponse dans la Survey of Income and Program Participation (SIPP), 3) une analyse du traitement de la non-réponse au questionnaire dans les recensements et enquêtes à caractère économique, et 4) une section concernant l'imputation pour les gains dans le cas de la Current Population Survey. En plus de présenter les diverses méthodes de compensation de la non-réponse utilisées par le Census Bureau, les auteurs examinent les problèmes particuliers inhérents à ces méthodes et décrivent les travaux et préoccupations courantes du Bureau relativement à la recherche sur la non-réponse.

2. LA NON-RÉPONSE DANS LES ENQUÊTES-ÉCHANTILLONS DÉMOGRAPHIQUES

Il peut arriver, un moment donné, que le Bureau of the Census s'occupe de près d'une trentaine d'enquêtes démographiques périodiques ou spéciales. Ces enquêtes portent sur l'activité, le revenu des particuliers et des familles, les soins médicaux, le transport, les loisirs, la criminalité et d'autres sujets qui reflètent les intérêts courants des citoyens, gouvernements, entreprises et institutions du pays. Les taux de non-réponse au questionnaire dans ces enquêtes varient de trois et quatre pour cent dans la National Crime Survey à plus de 25% (taux relevé pour la National Survey of Natural and Social Scientists and Engineers de 1984).

¹ David W. Chapman et Leroy Bailey, chercheurs principaux (Principal Researchers), Statistical Research Division, U.S. Bureau of the Census, Washington D.C. 20233. Daniel Kasprzyk, adjoint spécial, Office of the Chief, Population Division, U.S. Bureau of the Census, Washington D.C. 20233.

4. Utiliser une approche bivariable ou multivariée pour estimer de façon précise les points de retournement à la fin de la série. La relation avance-retard entre les demandes de prestations et les bénéficiaires peut contribuer à désaisonnaliser la série des bénéficiaires. Elle réduit la probabilité de considérer un point de retournement irrégulier comme un point de retournement cyclique. Comme l'avance est d'environ de cinq à six mois, le point de retournement de septembre 1982 dans la série des demandes de prestations confirme que le modèle multiplicatif appliqué à la série des bénéficiaires a indiqué un faux point de retournement en octobre 1982. Cependant, l'indicateur avancé prédit un point de retournement vers mars 1983 dans la série des bénéficiaires.
5. Utiliser l'option ARMMI avec des coefficients saisonniers coïncidants. Elle donne habituellement des révisions plus petites aux coefficients saisonniers lorsqu'une désaisonnalisation additive ou multiplicative est effectuée. Cependant, un faux point de retournement peut difficilement être corrigé par extrapolation, lorsque il est imputable au choix du mauvais modèle de décomposition.
6. Vérifier les données brutes et désaisonnalisées. On ne peut se fier à des tests uniques. Ainsi, l'ensemble de statistiques de contrôle de la qualité inclus dans le programme X-11-ARMMI n'est pas destiné à détecter une sous-estimation ou une sur-estimation de la série ou de faux points de retournement.
7. Toutes les recommandations ci-dessus sont valables si la série n'est pas affectée par les fluctuations des jours ouvrables. Si de telles fluctuations sont présentes, il faut les éliminer avant d'utiliser l'option ARMMI.

6. CONCLUSION

La désaisonnalisation d'une série temporelle n'est pas une procédure simple, en particulier lorsque le niveau d'une série a presque doublé en tout juste un an. La récession de 1981-82 a eu un fort impact très brusque non seulement sur la structure de la série, mais sur l'estimation de la tendance-cycle et de la composante saisonnière à la fin de la série. Par conséquent, il faut s'attendre à de sérieux problèmes de désaisonnalisation. L'analyse de la série des bénéficiaires a révélé que la sélection du mauvais modèle de décomposition se traduit par une sur-estimation considérable des chiffres désaisonnalisés pour les mois à saisonnalité faible. L'inverse est également vrai. De plus, on a relevé un faux point de retournement.

BIBLIOGRAPHIE

BOX, G.E.P., et JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

DAGUM, E.B. (1979). Data Extrapolation and smoothing with the X-11-ARIMA Seasonal Adjustment Method. *Proceedings of the 12th Annual Symposium Interface Computer Science and Statistics* (ed. Jane F. Gentleman), University of Waterloo, 195-202.

DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method*. Statistics Canada, Catalogue No. 12-564E.

HIGGINSON, J. (1977). *User Manual for the Decomposition Test*. Time Series Research and Analysis Division, Statistics Canada, Reference No. 77-01-001.

KLEIN, P.A., et MOORE, G.H. (1982). *The Leading Indicator Approach to Economic Forecasting Retrospect and Prospect*. Center for International Business Cycle Research, Rutgers University, Newark, N.J.

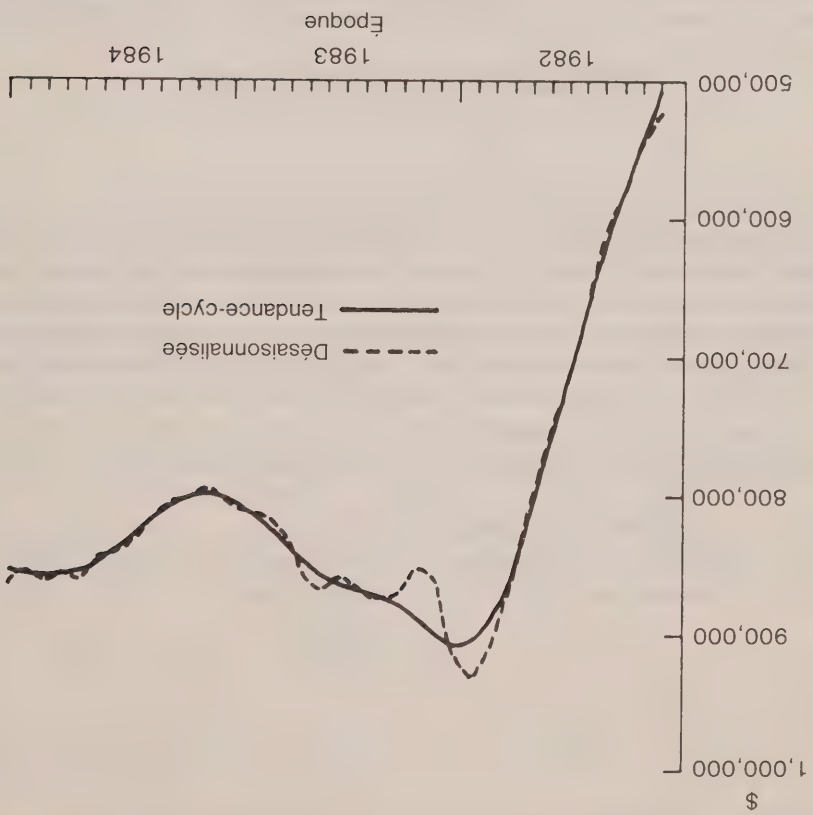


Figure 4. Prestations versées (Désaisonnalisation multiplicative).

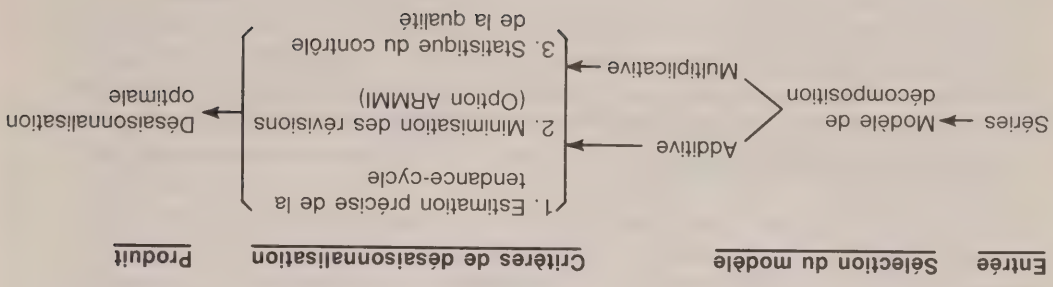


Figure 5. Procédures optimales de désaisonnalisation

La figure 3 présente la désaisonnalisation multiplicative de la série des bénéficiaires utilisant à la fois les extrapolations de la tendance à la hausse et de la tendance à la baisse. On peut constater, en la comparant à la figure 1, que les extrapolations ARMMI ne modifient pas les estimations multiplicatives de la tendance-cycle de l'année précédente. Le modèle multiplicatif indique toujours un point de retournement vers le mois d'octobre (tendance à la baisse, transformation log). Le modèle multiplicatif appliqué soit à la série des bénéficiaires non prolongée (figure 1), soit à la série prolongée, est donc douteux.

À la fin de 1983, on peut constater que le vrai point de retournement s'est produit en fait vers février 1983. Par conséquent, le point de retournement d'octobre ou de novembre 1982 peut difficilement être rectifié par extrapolation lorsque il est déterminé par un choix erroné du modèle de décomposition.

Le sur-ajustement et les problèmes d'identification des points de retournement se sont produits dans d'autres séries également. Ainsi, la figure 4 présente la série des "prestations versées" dans le cas d'une désaisonnalisation multiplicative avec des données réelles allant jusqu'à la fin de 1984. La série désaisonnalisée tend à osciller de façon systématique autour de la courbe de la tendance-cycle au point de retournement, surestimant et sous-estimant à la fois ainsi les prestations versées. Après le point de retournement, l'oscillation se fonde dans la courbe de la tendance-cycle, ce qui signifie que le modèle multiplicatif donne de mauvais résultats autour du point de retournement. Notons que cette série se caractérise par de fortes variations des jours ouvrables, qui ont été également supprimées.

5. SÉLECTION DE LA PROCÉDURE DE DÉSÉASONNALISATION OPTIMALE

La figure 5 résume les critères de désaisonnalisation qu'il faut prendre en compte pour régler les problèmes posés par l'interaction de la récession de 1981-82 et la désaisonnalisation des séries des bénéficiaires et des demandes de prestations. La sélection de la meilleure méthode de désaisonnalisation a été faite principalement à partir du premier critère.

Afin d'éviter la sur-estimation et la sous-estimation et de faux points de retournement dans les chiffres désaisonnalisés, il faut choisir le bon modèle de décomposition. Les données doivent faire l'objet d'une analyse méticuleuse, comme suit :

1. Effectuer un test de modèle sur les séries.
2. Désaisonnaliser la série à la fois de façon additive et multiplicative, si cela en vaut la peine. Si la différence entre les deux désaisonnalisations devient importante, comme à la figure 1, il faut vérifier la présence de sous-ajustement dans les mois de saisonnalité élevée et de sur-ajustement dans les mois de saisonnalité faible. Il faut également considérer le tableau D8 du programme X-11-ARMMI pour les tests F afin de vérifier la présence de saisonnalité stable et de saisonnalité mobile. Le modèle de décomposition qui ajuste mieux la série aura habituellement la valeur F plus élevée pour la saisonnalité stable, et la valeur F plus basse pour la saisonnalité mobile.
3. Vérifier les points de retournement. Dans le cas de la série des demandes de prestations, les deux modèles de décomposition ont indiqué un point de retournement en août ou septembre 1982. Par contre, dans le cas de la série des bénéficiaires, seul le modèle multiplicatif a indiqué un point de retournement en octobre 1982. Par conséquent, ou bien le modèle multiplicatif signale un faux point de retournement, ou bien le modèle additif manque ce dernier. L'analyse a révélé que ce point de retournement était faux, résultant d'une sur-estimation considérable du nombre désaisonnalisé de bénéficiaires dans les mois de saisonnalité faible, comme l'indique la figure 1.

Le programme automatique X-11-ARMMI se déroule comme suit:

- 1. Trois modèles univariés ARMMI de la forme multiplicative générale $(P,D,Q)_s$ (Box and Jenkins 1970) sont ajustés à la série mensuelle ou trimestrielle qui doit être désaisonnalisée. Les modèles sont

$$\begin{aligned} &(0,1,1)_s \quad (0,1,1)_s \\ &(0,2,2)_s \quad (0,1,1)_s \\ &(2,1,2)_s \quad (0,1,1)_s \end{aligned}$$

lorsque la série est désaisonnalisée de façon additive. Dans le cas d'une désaisonnalisation multiplicative, les mêmes modèles sont utilisés et les transformations log sont appliquées aux données pour les deux premiers modèles.

- 2. La série est extrapolée une année d'avance, et
- 3. à condition que les extrapolations soient acceptables, on applique alors la méthode ordinaire X-11 à la série ainsi prolongée.

La figure 3 présente la série des bénéficiaires désaisonnalisée à la fois de façon additive et de façon multiplicative avec les options automatiques X-11-ARMMI. Les modèles ARMMI qui ajustent et prédisent le mieux la série se terminant en décembre 1982 sont $(0,2,2)$ $(0,1,1)_{12}$, lorsque la série est désaisonnalisée de façon additive, et $\log (0,1,1)_{12}$ lorsqu'elle l'est de façon multiplicative. Le modèle $\log (0,2,2)$ $(0,1,1)_{12}$ a prévu une diminution de la série, tandis que le modèle $(0,2,2)$ $(0,1,1)_{12}$ a conservé la tendance à la hausse.

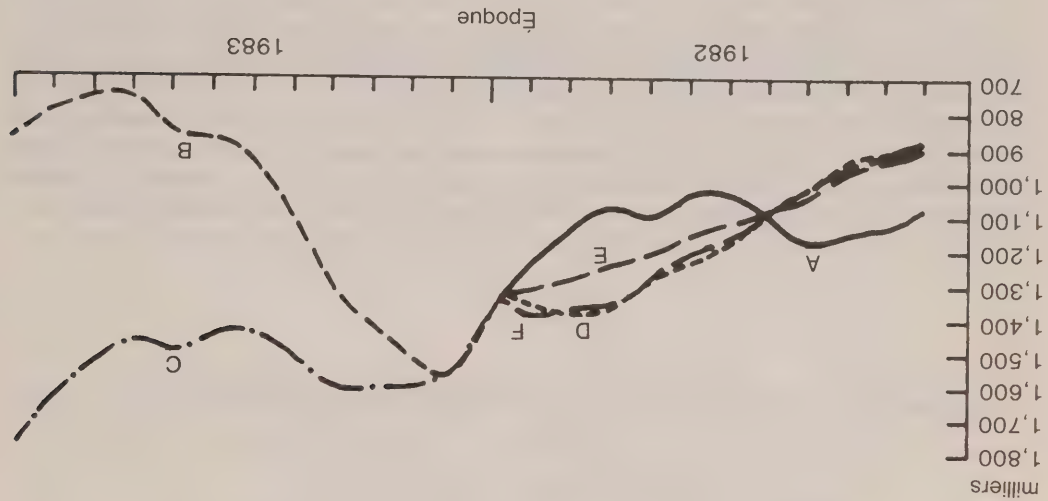


Figure 3. Série des bénéficiaires désaisonnalisée de façon additive et multiplicative avec différentes extrapolations ARMMI

- A. Série originale
- B. Extrapolations utilisant $\log (0,2,2)(0,1,1)$
- C. Extrapolations utilisant $(0,2,2)(0,1,1)$
- D. Désaisonnalisation additive utilisant des extrapolations ARMMI avec $\log (0,2,2)(0,1,1)$
- E. Désaisonnalisation multiplicative utilisant des extrapolations ARMMI avec $(0,2,2)(0,1,1)$
- F. Désaisonnalisation multiplicative utilisant des extrapolations ARMMI avec $\log (0,2,2)(0,1,1)$

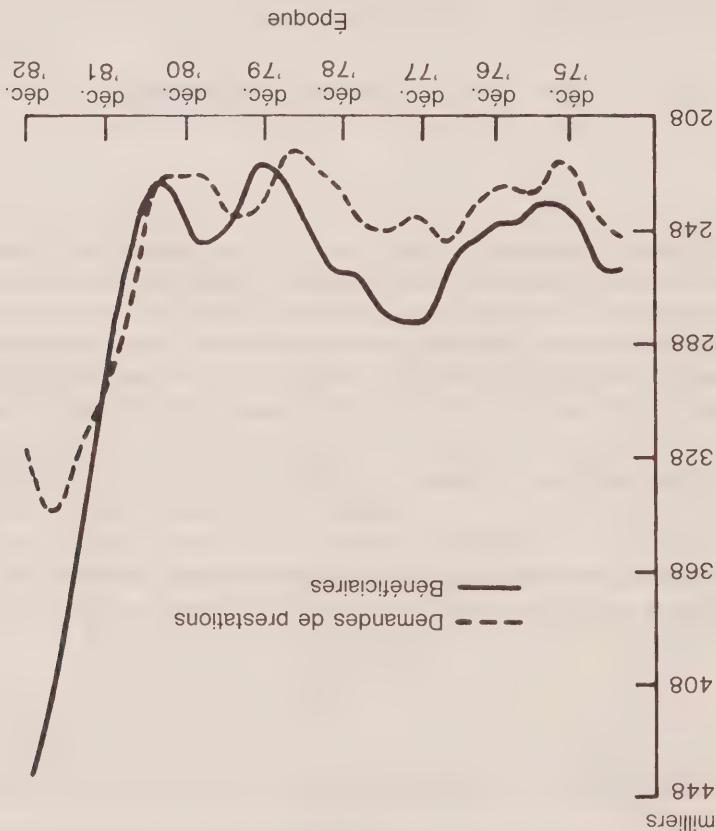


Figure 2. Demandes de prestations et bénéficiaires. Le nombre de bénéficiaires a été divisé par trois afin de rendre compatible l'échelle des deux séries.

La relation avance-retard entre les deux séries peut aider à la désaisonnalisation de la série des bénéficiaires. Elle réduit la probabilité de considérer par erreur un retournement irrégulier comme un point de retournement cyclique. La figure 2 montre que septembre 1982 est un point de retournement dans la série des demandes de prestations désaisonnalisée de façon multiplicative. C'est également le cas de la désaisonnalisation additive de la série. Comme les corrélations recoupées entre les deux séries font ressortir une relation avance-retard de cinq à six mois, le point de retournement de septembre 1982 de la série des demandes de prestations signifie que le modèle multiplicatif appliqué à la série des bénéficiaires a signalé un faux point de retournement vers octobre 1982. Cependant, l'indicateur avancé prédit un point de retournement vers mars 1983 dans la série des bénéficiaires.

4. EXTRAPOLATIONS ARMMI

Un procédé optimal de désaisonnalisation doit minimiser la révision des coefficients saisonniers courants et doit également donner des estimations fiables de la tendance-cycle, en particulier des points de retournement, à la fin de la série (Dagum 1979). L'analyse présentée dans les sections précédentes repose sur des données désaisonnalisées sans l'option ARMMI. Dans la présente section, nous allons nous attacher à l'utilisation des prévisions ARMMI, comme variable qui peut fournir une identification précise des points de retournement.

L'acceptation ou le rejet d'un modèle, dans le contexte d'une forte et brusque variation du niveau d'une série, doit visiblement reposer sur une analyse poussée des données. L'ensemble de statistiques du contrôle de la qualité inclus dans le programme X-11-ARMMI n'est pas destiné à détecter ce genre de problème dans le modèle. Dans cette expérience avec le modèle multiplicatif, aucune des dix statistiques de contrôle individuelles n'a fonctionné. Cependant, le test F de présence de saisonnalité mobile a révélé la présence d'une saisonnalité mobile croissante en 1982 dans les ratios finals SI non modifiés.

En plus d'un sur-ajustement et d'un sous-ajustement systématiques de la série, une autre conséquence de l'emploi d'un mauvais modèle de décomposition est la possibilité d'obtenir un point de retournement faux à la fin de la série.

Supposons qu'un point de retournement cyclique se produit si la série désaisonnalisée fait ressortir un changement de direction qui persiste pendant au moins cinq mois. Une fois que la série des bénéficiaires a été désaisonnalisée de façon multiplicative, la figure 1 révèle la présence possible d'un point de retournement vers le mois d'octobre 1982 lorsque la tendance à la hausse devient brusquement une tendance à la baisse. Ce point de retournement semble confirmé lorsque la série se terminant en décembre 1982 est prolongée d'un mois. La série désaisonnalisée de façon additive, par contre, ne fait ressortir aucun point de retournement. Les deux résultats sont donc contradictoires. Ainsi, ou bien le modèle multiplicatif indique un faux point de retournement, ou bien le modèle additif manque ce dernier.

Il n'est pas si facile de montrer que le modèle multiplicatif a signalé un faux point de retournement. Le modèle multiplicatif a créé un point de retournement vers octobre 1982. Le tableau 1 révèle qu'à très court terme, la mise à jour de la série n'efface pas ce point de retournement.

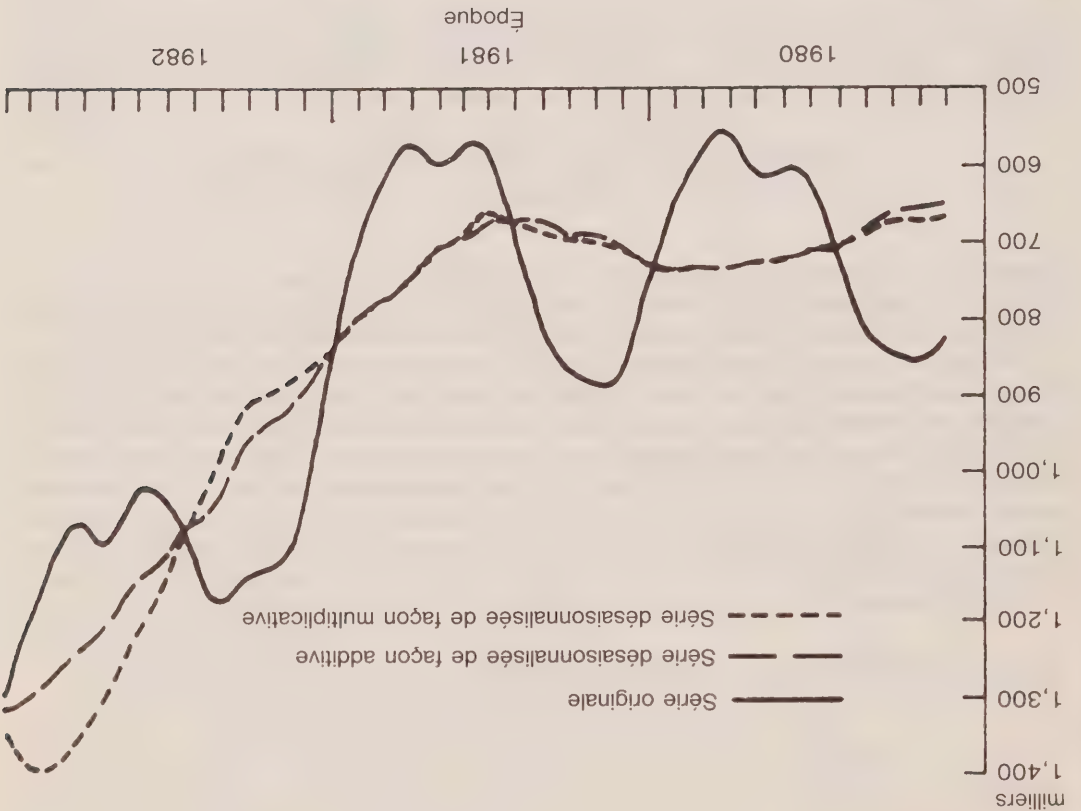
Tableau 1
Série des bénéficiaires désaisonnalisée de façon multiplicative
(milliers, juillet 1982 - février 1983)

Juillet	Août	Sept.	Oct.	Nov.	Déc.	Janv.	Fév.
124	131	140	142	141	131	121	123
124	130	140	141	141	131	121	
124	130	140	142	138	131		
124	130	140	142	141			
123	130	140	142	141			
123	129	139	142	141	134	121	

3. RELATION AVANCE-RETARD ENTRE LA SÉRIE DES DEMANDES DE PRESTATIONS ET LA SÉRIE DES BÉNÉFICIAIRES

Les indicateurs avancés sont sensibles à l'évolution de la conjoncture. Il s'agit de mesures des attentes ou des nouveaux engagements, et comme tels, ils donnent une indication préliminaire des changements attendus dans la tendance-cycle des indicateurs coïncidants et retardés.

La figure 2 montre la série des demandes de prestations comme un indicateur avancé de la série des bénéficiaires. La performance des indicateurs désaisonnalisés peut être testée à partir des critères de Klein et Moore (1982). Les deux séries satisfont à ces critères. D'abord, la correspondance entre les séries est de un à un, c'est-à-dire que le nombre de cycles est le même dans chaque série. Ensuite, il y a uniformité de période, les demandes de prestations étant toujours en avance. Enfin, il s'agit de séries mensuelles qui sont courantes, ou variation à la hausse ou à la baisse de la tendance de la série des bénéficiaires.



Le modèle multiplicatif suppose que la variation saisonnière est proportionnelle au niveau de cette dernière entre juin et novembre, qui sont les mois à saisonnalité faible. Comme le montre la figure 1, en sous-estimant le nombre de bénéficiaires saisonniers, le modèle multiplicatif a considérablement surestimé le nombre de bénéficiaires désaisonnalisés. L'in-

verse est également vrai.

Le modèle additif, par contre, ne suppose pas que les composantes de la série évoluent de façon proportionnelle. La figure 1 confirme que la tendance-cycle a augmenté, tandis que l'amplitude saisonnière est restée constante. Le modèle additif fournit par conséquent la meilleure désaisonnalisation. Il a donné de meilleurs résultats en 1982 que le modèle multiplicatif, et il est acceptable en 1980 et en 1981.

Au milieu de l'année 1982, il n'était pas facile de dire lequel des deux modèles, l'additif ou le multiplicatif, ajusterait le mieux la série des bénéficiaires. Comme cette série a été ajustée de façon multiplicative jusqu'en juin 1981, on pourrait normalement continuer de le faire en 1982. En 1982, y a-t-il eu des indices ou des preuves révélant que le modèle multiplicatif ne convenait plus?

Les prévisions ARMMI aident en général à réduire la révision des coefficients saisonniers et peuvent aider à mieux identifier les points de retournement à la fin de la série. Cette question sera abordée à la section 4.

2. MODÈLES DE DÉCOMPOSITION POUR LA DÉSAISONNALISATION

La plupart des séries des demandes de prestations et des bénéficiaires ont des caractéristiques semblables, aussi avons-nous décidé d'étudier une série de demandes de prestations et une série de bénéficiaires qui peuvent clairement illustrer quelques-uns des problèmes propres à la désaisonnalisation en récession grave. Il convient de noter que les résultats de notre analyse sont également valides pour une brusque et forte expansion au sein de l'économie. C'est la brusque et importante variation du niveau de la série entraînée par la récession ou par l'expansion qui est importante.

Le programme X-II-ARMMI (Dagum 1980) servira à la désaisonnalisation de ces séries. Le programme est appliqué à la série des demandes de prestations et à celle des bénéficiaires, et utilise les données à partir de janvier 1973 et mai 1975 respectivement pour l'estimation des composantes des séries temporelles. Le programme suppose une relation additive entre les composantes

$$O_t = TC_t + S_t + I_t \tag{2.1}$$

ou multiplicative

$$O_t = TC_t I_t \tag{2.2}$$

ou log additive

$$\log O_t = \log TC_t + \log S_t + \log I_t \tag{2.3}$$

où O désigne la série observée et brute; TC est la tendance-cycle, S et I sont les composantes saisonnière et irrégulière et t , le temps.

La désaisonnalisation consiste à supprimer la variation saisonnière S_t des données brutes O_t , ce qui donne une série désaisonnalisée, consistant en TC_t et I_t . Afin de savoir si une série multiplicative donne le meilleur ajustement, il est possible d'effectuer un test de présence de saisonnalité et un test de modèle sur les séries (Higginson, 1977). Le premier test révèle que les deux séries contiennent une très forte saisonnalité. Selon le deuxième test, le modèle multiplicatif ajuste mieux la série des bénéficiaires, lorsqu'on le teste pour la période mai 1975 - juin 1981. Lorsque la série est prolongée jusqu'en février 1983, en prenant en compte l'impact de la récession, le modèle additif donne un meilleur ajustement. Par contre, le test du modèle ne favorise ni le modèle additif, ni le modèle multiplicatif, pour la série des demandes de prestations.

On ajuste habituellement la série en utilisant un seul modèle, mais la figure 1 présente la série des bénéficiaires, corrigée par les deux modèles, mais sans l'option ARMMI. En 1980 et 1981, la différence entre les corrections additive et multiplicative a été petite par rapport à la différence observée en 1982.

La désaisonnalisation additive et la désaisonnalisation multiplicative en présence de variations rapides de la tendance-cycle¹

GUY HUOT et NAZIRA GAIT²

RÉSUMÉ

La désaisonnalisation d'une série temporelle n'est pas une procédure simple, en particulier lorsque le niveau d'une série a presque doublé en un an. La récession de 1981-82 a eu un fort impact très brusque non seulement sur la structure des séries, mais également sur l'estimation de la tendance-cycle et de la composante saisonnière à la fin de la série. Des problèmes de désaisonnalisation sérieux peuvent se poser. On peut citer comme exemple la sélection du mauvais modèle de décomposition, qui peut se traduire par un sous-ajustement des mois à saisonnalité élevée et par un sur-ajustement dans le cas des mois à saisonnalité faible. Ce modèle peut également donner un faux point de retournement. Les auteurs analysent ces deux aspects de l'interaction entre une récession grave et la désaisonnalisation.

MOTS CLÉS: Modèles de décomposition; ARMMI; relations avance-retard.

1. INTRODUCTION

Les années 1981 et 1982 ont été des années atypiques affectées par une grave récession. Cette dernière a eu un effet profond sur l'évolution et la structure des séries temporelles économiques et par conséquent sur leur désaisonnalisation. Les séries désaisonnalisées sont nécessaires pour poser le diagnostic de la santé socio-économique d'un pays. À leur tour, les politiques économiques et sociales qui reposent sur ces données se répercutent sur les décisions dans les secteurs privé et public. Cette récession par conséquent soulève un grand nombre de questions. On peut en conclure très vite qu'un examen de la désaisonnalisation s'impose. Les séries retenues ici sont: les demandes de prestations initiales et renouvelées reçues (prestations d'assurance-chômage) et les bénéficiaires. Il est difficile de voir comment leur tendance et leur cycle évoluent lorsque ces séries sont contaminées par la saisonnalité, à savoir, ici, les facteurs climatiques et institutionnels infra-annuels. La désaisonnalisation permet de mieux détecter les tendances fondamentales telles que les points de retournement et l'évolution de l'état actuel de l'économie.

On se propose dans cette communication d'analyser certains aspects de l'interaction entre une récession grave et la désaisonnalisation. En une seule année, en 1981, cette récession s'est traduite par pratiquement le doublement du nombre de bénéficiaires. Une forte augmentation aussi brusque soulève des questions à propos de la structure de la série, le choix du modèle de décomposition X-11-ARMMI, la détermination des points de retournement à la fin de la série et l'utilisation des prévisions ARMMI pour la désaisonnalisation. Dans la section 2, on examine deux importantes conséquences de l'emploi d'un mauvais modèle de décomposition, à savoir un sur-ajustement et un sous-ajustement systématiques des séries et la possibilité d'obtenir un faux point de retournement à la fin de la série. À la section 3, on utilise la relation avance-retard entre les séries des demandes de prestations et des bénéficiaires pour aider à la désaisonnalisation de cette dernière série.

¹ Cette communication a été présentée à la 145^e réunion annuelle de l'American Statistical Association, Las Vegas, Nevada, 1985.
² Guy Huot, Division des séries chronologiques recherche et analyse, Statistique Canada. N. Gait, Université de Sao Paulo, Brésil, en visite à Statistique Canada au moment de la rédaction.

REMERCIEMENTS

Les auteurs tiennent à remercier les adjoins à la rédaction pour les commentaires constructifs qu'ils ont formulés sur des versions antérieures de cet article.

BIBLIOGRAPHIE

DOLSON, D., GILES, P., et MORIN, J.-P. (1984). Méthode d'enquête sur les personnes souffrant d'une incapacité à l'aide de questions supplémentaires de l'Enquête sur la population active. *Techniques d'enquête*, 10, 203-214.

LAZARUS, G. (1985a). Characteristics of potentially disabled individuals based on the cluster analysis of activities of daily living. Document de travail, Division des méthodes d'enquêtes-institutions et agriculture. Statistique Canada.

LAZARUS, G. (1985b). An application of the results of the cluster analysis of activities of daily living. Document de travail, Division des méthodes d'enquêtes-institutions et agriculture. Statistique Canada.

RAYMOND, L., CHRISTE, E., et CLEMENCE, A. (1981). Vers l'établissement d'un score global d'incapacité fonctionnelle sur la base des questions de l'OCDE, d'après une enquête en Suisse. *Revue d'épidémiologie et santé publique*, 29, 451-459.

5.2.3 Echelles

Le tableau 6 donne l'ordre de classement des grappes selon les scores obtenus aux quatre premières composantes principales et le facteur $E(NADL)$. Il convient ici de rappeler que les coefficients de saturation des composantes sont fondés sur 11,412 cas et sur les réponses fournies aussi bien dans le questionnaire supplémentaire que dans le questionnaire de sélection, tandis que le facteur $E(NADL)$ est fondé sur 12,907 cas et sur les réponses fournies dans le questionnaire de sélection uniquement.

Le classement des grappes selon les composantes principales s'est effectué de la façon suivante. En ce qui concerne la composante représentant un indice global de l'état de santé (GLOBAL), les grappes ont été classées par ordre décroissant selon leur score. En ce qui concerne la deuxième composante principale (AHV/M), les grappes caractérisées par des troubles de mobilité se trouvent surtout vers le bas de l'échelle tandis que les grappes caractérisées par des troubles d'agilité ou des troubles d'agilité, des troubles auditifs ou des troubles de la vue se trouvent surtout vers le haut de l'échelle. En ce qui a trait à la troisième composante principale (MH/AV), les grappes caractérisées par des troubles de mobilité ou des troubles auditifs se trouvent surtout vers le bas de l'échelle tandis que les grappes caractérisées par des troubles d'agilité ou des troubles de la vue figurent surtout vers le haut de l'échelle. Enfin, les grappes caractérisées par des troubles d'agilité viennent avant les autres sur l'échelle MV/A de la quatrième composante principale. Etant donné la bipolarité des trois dernières composantes, nous avons dû déterminer arbitrairement une base pour l'échelle d'incapacité. Comme la grappe 8 présentait un facteur $E(NADL)$ très élevé, nous avons décidé de placer cette grappe à la tête des quatre autres échelles.

Le rang de la plupart des grappes varie beaucoup selon les échelles. Ces variations reflètent la nature des critères en fonction desquels les échelles ont été définies. La première composante principale, qui offre un indice global de l'état de santé, est peut-être le mode de classement des grappes le plus approprié. Premièrement, cette composante englobe les données qui ont été recueillies à l'aide du questionnaire de sélection et qui ont servi à la détermination du facteur $E(NADL)$. C'est pourquoi les échelles GLOBAL et $E(NADL)$ présentent à peu près le même ordre de classement. Le fait que cette composante englobe aussi les données recueillies au moyen du questionnaire supplémentaire nous porte à croire que l'échelle GLOBAL est supérieure à des échelles telles que $E(NADL)$. Pour le bénéfice du lecteur, signalons que les vingt-neuf grappes ont toutes été classées et que leur classement initial dans le tableau 6 a été établi en fonction des groupes primaires. Les données relatives à ces groupes n'ont toutefois pas servi à l'analyse en composantes principales.

6. CONCLUSIONS

Suivant les réponses fournies dans le questionnaire de sélection de l'enquête, les personnes visées ont été classées par grappe selon des caractéristiques communes. Les grappes ont ensuite été classées en fonction des données contenues dans le questionnaire de sélection (classement incomplet fondé sur la valeur de $E(NADL)$ et présenté dans le tableau 4) et, en dernier lieu, en fonction des données contenues dans le questionnaire de sélection et le questionnaire supplémentaire (échelle GLOBAL présentée dans le tableau 6). On estime actuellement que l'échelle GLOBAL représente le mode de classement le plus approprié parmi toutes les échelles analysées dans cette étude. Nous pouvons par ailleurs prétendre qu'il est impossible, à l'heure actuelle, de définir un indice unique du degré d'incapacité; en fait, cet indice devrait être quadridimensionnel pour tenir compte des quatre composantes principales.

Agilité présente un écart négatif. Cependant, le résultat observé sur ce dernier groupe est inclus parmi les troubles d'agilité. Or, les personnes ayant des problèmes avec l'A.Q. A26 font partie du groupe souffrant d'une incapacité particulière, le groupe présentant un écart clairement négatif pour la première composante principale.

ii) La deuxième composante principale établit une nette distinction entre les troubles de mobilité (–) d'une part et les troubles d'agilité, les troubles auditifs et les troubles de la vue (+) d'autre part. Comme on pouvait s'y attendre, des écarts positifs sont enregistrés pour les groupes primaires Ouïe/Vue, Ouïe, Vue et Agilité, tandis que des écarts négatifs sont observés pour les groupes Mobilité/Agilité, Mobilité et Aucune de ces incapacités. L'écart enregistré pour le groupe Incapacité particulière est presque nul.

iii) La troisième composante principale établit une nette distinction entre les troubles de mobilité et les troubles auditifs (+) d'une part, et les troubles d'agilité et ceux de la vue (–) d'autre part. Cette fois encore, les résultats sont cohérents.

iv) La quatrième composante principale établit une nette distinction entre les troubles de mobilité et les troubles de la vue (+) d'une part, et les troubles d'agilité (–) d'autre part. Une fois de plus, les résultats observés sont conformes à la construction des groupes primaires.

Tableau 6

Classement des grappes selon différentes échelles

Grappe	Sigle	PRIN1 (Global)	PRIN2 (AHV/M)	PRIN3 (MH/AV)	PRIN4 (MV/A)	E(NADL)
2	HVMA1	9	4	27	28	5
5	HVNI	22	2	22	25	12
1	HMA1	3	3	24	6	1
3	HMA2	10	14	28	10	7
4	HMI	16	15	29	20	13
6	HA1	20	8	25	3	15
7	HNI	29	7	26	9	24
9	VM1	2	6	4	23	3
12	VM2	4	10	7	27	6
13	VM1	13	11	11	29	11
21	VNI	23	5	2	26	20
8	MA1	1	1	1	1	2
10	MA2	5	20	13	4	4
14	MA3	6	24	16	7	8
11	MA4	7	23	17	8	9
15	MA5	8	28	20	18	10
18	M1	14	26	19	21	14
16	M2	15	25	18	17	18
19	M3	11	29	23	24	19
20	M4	18	27	21	22	22
22	A1	17	9	3	2	17
23	N1	21	17	6	5	21
25	N2	19	22	10	16	23
27	N3	24	19	10	12	26
28	N4	28	12	15	11	28
29	N5	25	16	12	15	27
26	N6	26	18	8	14	28
17	SMA1	12	21	14	19	16
24	SNI	27	13	5	13	29

La deuxième composante principale présente des coefficients de saturation négatifs pour les questions A10, A11, A12, A14 et A15. Toutefois, le coefficient pour A15 est presque nul. Cette composante présente des coefficients positifs pour les A.Q. qui ont trait à des troubles d'agilité, de même que pour les A.Q. qui ont trait à des troubles de la vue. La deuxième composante semble donc établir une distinction nette entre, d'une part, les troubles d'agilité, les troubles auditifs et les troubles de la vue et, d'autre part, les troubles de mobilité. Elle est désignée par le sigle "AHV/M".

La troisième composante principale présente des coefficients de saturation positifs pour les A.Q. qui ont trait à des troubles de mobilité et à des troubles auditifs, et des coefficients négatifs pour les A.Q. qui se rattachent surtout à des troubles d'agilité ou des troubles de la vue. Elle est désignée par le sigle "MH/AV".

La quatrième composante principale présente des coefficients de saturation positifs pour les A.Q. qui ont trait à des troubles de mobilité ou des troubles de la vue et des coefficients négatifs pour les A.Q. qui se rattachent surtout à des troubles d'agilité. Cette composante est désignée par le sigle "MV/A".

5.2.2 Scores moyens

Le tableau 5 donne les écarts moyens entre les scores obtenus aux différentes composantes principales pour chaque groupe primaire et les scores moyens des 11,412 unités de l'échantillon pour chaque groupe primaire. Nous sommes maintenant en mesure de vérifier si la classification incomplète proposée plus haut est conforme aux résultats de l'analyse en composantes principales. Il convient de préciser que les observations suivantes sont fondées sur les données du tableau 5.

i) En ce qui concerne la première composante principale (GLOBAL), l'écart le plus élevé est enregistré pour le groupe primaire Mobilité/Agilité tandis que l'écart le plus faible est enregistré pour le groupe primaire "Aucune de ces incapacités". Le groupe de personnes souffrant surtout de troubles de la vue ou de troubles auditifs est caractérisé par un écart moyen positif; c'est aussi le cas du groupe de personnes souffrant surtout de troubles de la vue. En revanche, le groupe de personnes souffrant surtout de troubles auditifs montre un écart négatif, ce qui signifie que les personnes qui ont de la difficulté à entendre ne souffrent généralement pas d'autres incapacités. On pourrait être porté à dire la même chose du groupe de personnes souffrant surtout de troubles d'agilité. On constate en effet que les groupes Mobilité/Agilité et Mobilité présentent chacun un écart positif, alors que le groupe

Tableau 5
Écarts moyens entre les scores de chaque groupe primaire et les scores moyens totaux obtenus sur chaque composante principale

Écarts				
Groupe primaire	Taille de l'échantillon	PRIN1 (Global)	PRIN2 (AVH/M)	PRIN3 (MH/AV)
Quie/Vue	346	0.68	1.26	0.61
Quie	2741	-0.33	0.54	0.81
Vue	888	0.30	0.69	-0.76
Incapacité particulière	151	-1.02	-0.04	-0.47
Mobilité/Agilité	1311	3.31	-0.33	-0.21
Mobilité	1893	0.30	-0.80	0.18
Agilité	195	-0.19	0.31	-0.80
Aucune de ces incapacités	3887	-1.11	-0.16	-0.41
				-0.22

5. CARACTÉRISTIQUES DES GRAPPES

Nous appliquons la méthode des composantes principales pour analyser les caractéristiques des grappes qui ont été formées. Raymond *et coll.* ont aussi utilisé cette méthode; dans notre cas, l'analyse repose sur des moyennes de groupes plutôt que sur des individus.

5.1 Méthode

On considère ici un sous-ensemble de personnes visées par l'enquête pour lesquelles des renseignements supplémentaires sont disponibles. Nous avons, notamment, tenu compte des réponses données à des questions de la forme: (B101) . . . Est-il (elle) incapable de marcher sur une distance de 400 mètres sans se reposer? Ce genre de question a été posé pour chacune des A.Q., de A10 à A26. Ainsi, des 12,907 personnes qui faisaient partie de l'échantillon initial, 11,412 ont pu être retenues pour notre analyse. Le reste (1,495 personnes) a été exclu de l'analyse à cause des problèmes de non-réponse. Lorsque la personne handicapée avait indiqué qu'elle était incapable d'accomplir une A.Q. donnée, on inscrivait un "1"; dans le cas contraire, on inscrivait un "0".

Les moyennes ont été calculées pour chacune des dix-neuf questions de sélection et des dix-sept questions supplémentaires, pour chaque grappe. Ensuite, les moyennes établies aux questions visant à savoir si une personne était incapable d'accomplir telle ou telle activité ont été multipliées par le rapport du nombre total moyen de "oui" parmi A10 à A26 et du nombre total moyen de cas d'incapacité parmi B101 à B261, afin d'obtenir des mesures cohérentes et d'éviter les problèmes de changement d'échelle qui peuvent se produire lors de l'analyse en composantes principales.

Nous avons déterminé les composantes principales en utilisant les moyennes calculées pour les dix-neuf questions de sélection et les dix-sept questions supplémentaires comme variables et les grappes comme observations, et en pondérant selon les tailles des grappes. Nous avons ensuite classé les grappes en fonction des scores obtenus à chacune des quatre premières composantes principales.

En dernier lieu, nous avons classé les éléments des grappes en fonction des groupes primaires définis précédemment et calculé les scores moyens obtenus aux quatre premières composantes principales pour chacun des huit groupes primaires, les facteurs de pondération étant en l'occurrence le nombre d'unités comprises dans ces groupes.

5.2 Résultats

Les résultats sont présentés en deux volets. Dans le premier volet, nous analysons les composantes principales et tentons de les désigner par un sigle selon les coefficients de saturation. Nous examinons aussi la composition des groupes primaires en fonction des moyennes des composantes principales. Dans le deuxième volet, nous étudions le classement des grappes selon les quatre premières composantes principales.

5.2.1 Composantes

Les quatre premières composantes principales pour les dix-neuf questions de sélection et les dix-sept questions supplémentaires ont expliqué un peu plus que sept huitièmes de la variance totale et semble tout à fait convenir aux objectifs de notre analyse.

Les coefficients de saturation de la première composante principale sont tous positifs sauf en ce qui a trait à quatre questions (A24, A25 et B241 ont trait à des troubles auditifs tandis que A28 concerne le handicap mental). Les coefficients négatifs sont proches de la valeur nulle. La première composante principale semble donc être un indice global de l'état de santé. Elle explique près de 66% de la variance totale et est désignée par le terme "GLOBAL".

Tableau 4
Classement des grappes par groupe primaire

Groupe primaire	Grappe	Taille de l'échantillon	E(NADL)	Stige
HV (Oùc/Vue)	2	187	8.566	HVMA1
	5	203	4.895	HVNI
	1	303	11.855	HMA1
	3	355	7.488	HMA2
H (Oùe)	4	311	4.829	HMI
	6	289	4.760	HA1
	7	1,770	2.120	HNI
	9	56	9.841	VMA1
V (Vue)	12	160	7.783	VMMA2
	13	164	4.976	VM1
	21	618	2.756	VNI
	17	24	4.708	SMA1
S (Incapacité particulière)	24	246	0.482	SNI
	8	245	11.480	MA1
	10	210	8.947	MA2
	11	166	6.819	MA4
MA (Mobilité/Agilité)	14	187	6.924	MA3
	15	677	6.759	MA5
	16	458	4.374	M2
	18	173	4.780	M1
M (Mobilité)	19	582	3.905	M3
	20	857	2.290	M4
	22	215	4.614	A1
	23	1,164	2.687	N1
N (Aucune de ces incapacités)	25	295	2.273	N2
	26	1,923	0.573	N6
	27	371	1.486	N3
	28	204	1.462	N4
A (Agilité)	29	494	1.190	N5

4.3 Degré d'incapacité

Une question à laquelle on s'intéresse sur le plan analytique est l'établissement d'un indice du degré d'incapacité. L'ouvrage de Raymond *et coll.* (1981) compte parmi ceux qui ont déjà examiné cette question.

Un indice du degré d'incapacité est utile dans la mesure où il permet d'établir des comparaisons simples entre les degrés d'incapacité dont souffrent les personnes visées par l'enquête. L'emploi du facteur $E(NADL)$ dans de telles comparaisons suppose que les genres d'incapacité sont autoperçus; on remarque, à ce propos, que deux des A.Q. ont trait à des troubles auditifs tandis que quatre se rapportent à des troubles de mobilité. Souignons de plus qu'un score unique comme le facteur $E(NADL)$ n'exprime pas le caractère multidimensionnel du degré d'incapacité.

Dans le tableau 4, les grappes sont classées suivant le degré d'incapacité de leurs unités respectives à l'intérieur de chaque groupe primaire. Ce classement intra-groupe est plus conforme à l'idée du caractère multidimensionnel du degré d'incapacité que pourrait l'être un classement général.

Tableau 3

Nombre moyen de cas d'incapacité selon le genre d'incapacité

Grappe	Ouïe	Vue	Mobilité	Agilité	Total
1	1.733	0.657	3.624	5.841	11.855
2	1.717	1.508	3.171	2.170	8.566
3	1.637	0.034	3.274	2.543	7.488
4	1.579	0.016	2.582	0.710	4.887
5	1.596	1.463	0.625	1.211	4.895
6	1.509	0.014	1.091	2.152	4.766
7	1.605	0.013	0.253	0.246	2.117
8	0.012	0.493	3.772	7.203	11.480
9	0.054	1.304	3.643	4.480	9.841
10	0.000	0.005	3.686	5.256	8.947
11	0.006	0.018	3.476	3.319	6.819
12	0.044	1.456	3.445	2.838	7.783
13	0.012	1.427	2.653	0.884	4.976
14	0.009	0.010	3.727	3.178	6.924
15	0.021	0.021	3.776	2.941	6.759
16	0.004	0.000	2.406	1.964	4.374
17	0.000	0.000	2.625	2.083	4.708
18	0.000	0.000	2.890	1.890	4.780
19	0.002	0.005	3.404	0.494	3.905
20	0.004	0.007	2.046	0.233	2.290
21	0.026	1.411	0.467	0.852	2.756
22	0.014	0.014	1.088	3.498	4.614
23	0.007	0.008	0.984	1.688	2.687
24	0.000	0.000	0.068	0.352	0.482
25	0.000	0.003	1.685	0.587	2.273
26	0.005	0.007	0.303	0.258	0.573
27	0.003	0.003	0.310	1.170	1.486
28	0.005	0.000	0.172	1.285	1.462
29	0.057	0.065	0.650	0.418	1.190

caractérisées par des troubles de mobilité ou d'agilité étaient désignées par la lettre "N". Ainsi, les sigles HMA1 et HMA2 désignent des grappes dont une forte proportion des éléments souffrent de troubles auditifs ou de troubles de mobilité ou d'agilité mais ne sont pas atteints de troubles de la vue. Par ailleurs, le sigle VNI désigne une grappe dont les éléments souffrent uniquement de troubles de la vue.

4.2 Groupes primaires

Les grappes qui partageaient sensiblement les mêmes handicaps ont été classées par groupes primaires, à l'intérieur desquels elles pouvaient être mieux comparées à l'aide du facteur $E(NADL)$. Le tableau 4 présente les grappes selon les groupes primaires auxquels elles appartiennent.

Tableau 2
Résultats de l'analyse typologique

Grappe	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
1	U	92.7	79.9	59.7	89.8	85.5	U	62.7	86.8	60.1
2	U	77.0	63.1	16.0	77.0	55.6	Z	11.2	46.5	31.0
3	U	85.1	66.5	19.4	75.8	U	Z	15.8	49.6	26.5
4	U	65.6	36.7	6.4	55.9	Z	Z	4.5	21.5	17.7
5	Z	18.7	18.2	3.4	25.6	24.6	4.9	6.9	21.7	20.7
6	Z	36.3	23.2	4.8	49.5	U	11.8	16.6	28.4	21.1
7	Z	9.2	5.3	0.3	10.8	Z	1.1	0.9	4.4	7.1
8	U	94.7	88.6	67.3	93.9	89.0	U	74.7	94.7	84.0
9	U	92.9	82.1	55.4	89.3	91.1	U	58.9	87.5	30.4
10	U	95.7	81.0	55.7	91.9	93.8	U	U	85.2	33.3
11	U	92.2	71.7	21.1	83.7	74.1	U	Z	59.0	28.9
12	U	91.9	71.3	25.0	81.3	U	Z	16.9	58.1	31.9
13	U	61.0	48.8	4.3	55.5	Z	Z	4.9	32.3	14.0
14	U	91.3	U	23.6	81.4	U	Z	16.8	40.2	U
15	U	93.6	U	29.9	84.0	U	Z	19.3	56.1	Z
16	U	74.9	Z	10.9	65.7	U	Z	12.9	32.8	16.4
17	U	66.7	58.3	12.5	37.5	Z	Z	0.0	37.5	20.8
18	U	74.0	55.5	7.5	59.5	Z	Z	10.4	29.5	U
19	U	79.6	U	11.5	60.8	Z	Z	2.9	14.6	Z
20	U	59.0	Z	2.7	45.6	Z	Z	2.2	10.4	Z
21	Z	14.7	12.6	1.9	19.4	13.9	5.5	4.7	22.2	11.5
22	Z	26.5	40.9	7.0	41.4	59.1	U	32.1	47.4	35.8
23	Z	29.0	26.1	2.1	43.3	U	Z	13.0	19.0	13.5
24	Z	2.4	2.4	0.0	2.0	Z	Z	0.4	7.7	3.3
25	Z	35.6	U	2.4	32.9	Z	Z	3.1	8.5	18.0
26	Z	13.5	Z	0.3	16.8	Z	Z	1.8	4.2	9.1
27	Z	17.0	13.7	0.3	U	Z	Z	0.1	7.8	U
28	Z	10.3	6.9	0.0	Z	Z	Z	0.1	7.8	U
29	Z	38.7	26.3	0.6	Z	Z	Z	2.2	10.9	Z
1	63.7	42.2	38.6	27.1	73.3	U	23.4	94.4	6.3	303
2	35.3	11.8	U	50.8	71.7	U	9.6	85.0	1.6	187
3	34.6	5.9	Z	3.4	63.7	U	2.5	88.7	1.1	355
4	16.4	1.9	Z	1.6	57.9	U	2.6	73.3	1.0	311
5	17.7	8.4	U	46.3	59.6	U	12.8	55.7	7.9	203
6	24.9	3.5	Z	1.4	50.9	U	4.2	71.3	1.0	289
7	4.6	0.6	Z	1.3	60.5	U	5.6	26.3	1.6	1,770
8	78.4	U	32.6	16.7	1.2	Z	32.2	96.3	9.8	245
9	50.0	Z	U	30.4	5.4	Z	10.7	100.0	5.4	56
10	55.2	Z	Z	0.5	0.0	Z	2.4	89.0	1.9	210
11	45.8	Z	Z	1.8	0.6	Z	3.0	90.4	0.6	166
12	39.4	7.5	U	45.6	4.4	Z	5.0	93.1	1.9	160
13	20.7	5.5	U	42.7	1.2	Z	6.7	78.0	4.3	164
14	34.4	1.5	Z	1.0	0.9	Z	1.3	89.4	1.2	187
15	66.3	16.6	Z	2.1	2.1	Z	5.9	92.0	1.6	677
16	20.7	0.7	Z	0.0	0.4	Z	2.0	82.3	0.4	458
17	16.7	20.8	Z	0.0	0.0	Z	U	91.7	33.3	24
18	29.5	12.1	Z	0.0	0.0	Z	Z	82.1	1.2	173
19	19.4	1.0	Z	0.8	0.0	Z	Z	73.5	1.0	582
20	8.0	0.0	Z	0.7	0.4	Z	Z	66.7	0.6	857
21	9.7	7.1	U	41.1	2.6	Z	8.7	55.3	9.2	618
22	41.9	19.5	Z	1.4	1.4	Z	7.0	76.3	4.7	215
23	18.1	1.9	Z	0.8	0.7	Z	1.2	66.6	0.4	1,164
24	0.8	2.0	Z	0.0	0.0	Z	27.2	62.2	U	246
25	23.7	1.4	Z	0.3	0.0	Z	1.4	U	Z	295
26	7.3	1.2	Z	0.7	0.5	Z	1.9	U	Z	1,293
27	2.4	0.3	Z	0.3	0.3	Z	0.3	Z	Z	371
28	11.8	8.3	Z	0.0	0.5	Z	0.5	Z	Z	204
29	18.0	1.6	Z	6.5	5.7	Z	8.5	Z	Z	494

En ce qui concerne l'étape du regroupement, on a appliqué des critères subjectifs en se fondant sur la grandeur de la valeur F , la taille des groupes et la dispersion des observations. La combinaison des groupes s'est effectuée, en majeure partie, dans l'ordre inverse de celui de la répartition.

Dans la première étape, on s'est servi de données fondées sur des covariances non pondérées, puis de données fondées sur des covariances pondérées. On a observé que les résultats étaient essentiellement les mêmes dans les deux cas. On a donc décidé, lors de cette étape, de mettre de côté les poids d'échantillonnage de façon à ne pas accroître inutilement la complexité du problème. De plus, on a estimé que les poids ne constituaient pas un élément important quant aux caractéristiques des unités en grappes; ils sont toutefois essentiels à l'évaluation et à l'analyse.

3.2 Description

L'analyse typologique a permis de grouper les personnes handicapées en fonction de profils comparables mais pas nécessairement identiques. Pour les besoins de notre étude, le profil d'un répondant était établi à partir des réponses fournies par cette personne aux 17 questions relatives aux activités quotidiennes (oui, éprouve des difficultés/non, n'éprouve pas de difficultés), de la réponse fournie à la question portant sur la restriction de mouvement dans certaines activités (oui/non) et de la réponse à la question sur l'existence d'un handicap mental.

Le tableau 2 expose en détail les grappes définitives. Les symboles "U" et "Z" indiquent comment ces groupes sont définis. Le symbole "U" signifie que, par définition, toutes les unités du groupe en question répondent oui à la question correspondante. Le symbole "Z" signifie que toutes les unités du groupe en question répondent non à la question correspondante. Il convient de souligner que six des dix-neuf questions de sélection n'ont pas servi explicitement à la classification des répondants. Ce sont les questions A11, A13, A18, A20, A23 et A24.

4. CARACTÉRISATION DES GRAPPES

Cette section se propose d'identifier les différentes grappes par sigle. On y présente les notions de "genre d'incapacité" et de "groupe primaire", et on y ordonne les grappes selon le degré d'incapacité.

4.1 Genre d'incapacité

Des valeurs critiques ont été déterminées pour faciliter le classement des grappes. Nous avons établi ces valeurs en classant les grappes selon le genre d'incapacité et en identifiant une valeur aberrante de $E(NADL)$ pour chaque genre d'incapacité, $E(NADL)$ étant défini comme le nombre moyen de cas d'incapacité relevés parmi les dix-sept activités quotidiennes (A10-A26). En règle générale, on pouvait affirmer qu'une grappe était caractérisée par un certain genre d'incapacité ou souffrait d'un certain genre d'incapacité lorsque la valeur de $E(NADL)$ pour ce genre d'incapacité dépassait la valeur critique établie. Par exemple, dans le cas des troubles de mobilité, on a établi une valeur de $E(NADL)$ pour les activités A10, A11, A12 et A14. Le tableau 3 donne les valeurs de $E(NADL)$ pour chaque grappe et pour chaque genre d'incapacité.

Les grappes ont été désignées par sigle de la façon suivante. Lorsqu'une grappe souffrait d'une incapacité quelconque, on incluait la lettre correspondante dans le sigle d'identification. Deux grappes qui renfermaient des unités ayant des troubles d'élocution ou souffrant d'un handicap mental ont été désignées comme "spéciales". Les grappes qui n'étaient pas

Les A.Q. sont énumérées dans le tableau 1, où l'on trouve également le numéro des questions correspondantes ainsi que les catégories auxquelles ces activités appartiennent. Deux A.Q. ont trait à des troubles auditifs, deux à des troubles de la vue, quatre à des troubles de mobilité, une à des troubles d'élocution et les huit autres à des troubles d'agilité. Les questions de sélection relatives aux activités quotidiennes étaient libellées de la façon suivante: (A20), par exemple, " . . . éprouve-t-il (elle) des difficultés à étendre le bras pour prendre quelque chose?" La question relative aux limitations d'activité (A27) demandait si la personne était limitée dans le genre ou la quantité d'activités qu'elle pouvait faire à la maison, au travail ou à l'école à cause d'une affection ou d'un problème de santé chronique. La dernière question de sélection (A28) demandait si la personne souffrait d'un handicap mental. Il convient de souligner que l'enquête visait seulement les personnes souffrant d'une affection ou d'un problème de santé chronique, c'est-à-dire qui dure depuis plus de six mois ou est censé durer plus de six mois (à l'exclusion d'une grossesse). Une personne était considérée comme invalide si elle éprouvait des difficultés à accomplir au moins une des activités quotidiennes, si elle était limitée dans une activité quelconque ou si elle souffrait d'un handicap mental. (Dans ce dernier cas, il fallait une réponse par personne interposée.)

2.1.2 Questions supplémentaires

Les questions supplémentaires étaient réservées aux personnes sélectionnées par le questionnaire de sélection. Une des questions supplémentaires visait à déterminer si la personne en question était totalement incapable d'accomplir la ou les A.Q. qu'elle avait indiquées à l'étape de la sélection. Les autres questions supplémentaires concernaient les points suivants: nature de l'incapacité (troubles de la vue, troubles auditifs, troubles d'élocution ou de mobilité); problèmes liés à l'aptitude au travail ou au lieu de travail; obstacles à la formation et disponibilité de services éducatifs; problèmes liés au déplacement sur de courtes ou de longues distances et problèmes liés au domicile de la personne handicapée et aux installations spéciales. À l'aide des réponses à ces questions, il était possible d'analyser les caractéristiques des grappes ou de construire un indice du degré d'incapacité (voir Lazarus 1985a, 1985b).

3. GRAPPES

Dans cette section, nous décrivons la méthode de formation des grappes. Cette méthode a été conçue spécialement pour le cas qui nous occupe. Les sections 3.2 et 3.3 fournissent les détails techniques relatifs à cette méthode. Tous les calculs ont été effectués à l'aide du SAS.

3.1 Méthode

Nous présentons ici un résumé de la méthode de formation des grappes définitives.

- a) l'étape de la répartition, où les 12,907 personnes visées par l'enquête ont été classées successivement à l'aide de PROC CANDISC;
- b) l'étape du regroupement, où les classes ont été combinées.

En ce qui concerne la première étape, la procédure décrite ci-dessous a été appliquée itérativement. On a tout d'abord regroupé toutes les observations en une seule grappe. À chaque étape, chacune des grappes existantes a été divisée en deux. Chacune de ces grappes a fait l'objet d'une analyse canonique, où les variables non constantes ont servi à tout de rôle de variable de regroupement tandis que toutes les autres ont servi de variables explicatives. La grappe a ensuite été divisée en deux suivant les résultats de l'analyse discriminante ayant produit la valeur F la plus élevée. On s'est trouvé ainsi à maximiser le déterminant de la matrice des sommes des carrés entre les groupes.

des six provinces les plus faiblement échantillonnées en octobre (c.-à-d. Terre-Neuve, Ile-du-Prince-Edouard, Nouvelle-Ecosse, Nouveau-Brunswick, Manitoba et Saskatchewan). Les enfants de toutes les provinces étaient visés lors des deux occasions d'enquête. Cet exposé porte essentiellement sur les données qui ont été tirées du questionnaire des adultes de l'enquête menée en octobre 1983. Cette enquête a permis de recueillir les réponses de 92,945 adultes, répartis dans environ 47,000 ménages.

2.1 Questionnaire

2.1.1 Questions de sélection

L'Enquête sur la santé et l'incapacité au Canada comportait un questionnaire de sélection visant à former un échantillon de personnes aptes à répondre à un deuxième questionnaire. Le questionnaire de sélection comportait dix-neuf questions, dont dix-sept portaient sur des activités quotidiennes, une sur les limitations d'activités et la dernière sur l'existence d'un handicap mental. Les activités quotidiennes (A.Q.) sont un ensemble d'activités qu'une personne accomplit dans la vie de tous les jours. La série d'activités utilisée ici est une variante de celle qui a été établie par l'Organisation de coopération et de développement économique (OCDE) et est utilisée dans plusieurs autres pays.

Tableau 1

Activités quotidiennes

N° de question	Description	Catégorie
A10	Marcher sur une distance de 400 mètres	Mobilité
A11	Monter et descendre un escalier	Mobilité
A12	Transporter un objet de 5 kg sur 10 mètres	Mobilité
A13	Se déplacer d'une pièce à une autre	Agilité
A14	Se tenir debout pendant de longues périodes	Mobilité
A15	En position debout, se pencher pour ramasser un objet	Agilité
A16	S'habiller et se déshabiller	Agilité
A17	Se mettre au lit et sortir du lit	Agilité
A18	Se couper les ongles d'orteils	Agilité
A19	Se servir de ses doigts pour saisir	Agilité
A20	Etendre le bras pour prendre quelque chose ou manier un objet	Agilité
A21	Couper ses aliments	Agilité
A22	Lire les caractères d'un journal	Vue
A23	Voir clairement la figure de quelqu'un	Vue
A24	Entendre ce qui se dit au cours d'une conversation avec une autre personne	Ouïe
A25	Entendre ce qui se dit au cours d'une conversation avec au moins deux autres personnes	Ouïe
A26	Parler et être compris	Elocution

Analyse typologique des activités de la vie quotidienne à partir de l'Enquête sur la santé et l'incapacité au Canada¹

D.A. BINDER et G. LAZARUS²

RÉSUMÉ

L'Enquête sur la santé et l'incapacité au Canada, enquête supplémentaire à l'Enquête sur la population active du Canada en octobre 1983, a permis de recueillir des données sur les personnes invalides à l'aide d'un questionnaire de sélection et d'un questionnaire supplémentaire remis uniquement aux personnes sélectionnées. Les données tirées du questionnaire de sélection, qui couvrait un ensemble d'activités quotidiennes, ont servi à grouper les répondants selon des caractéristiques identifiées. Les auteurs décrivent les groupes de répondants et analysent les méthodes utilisées pour former ces groupes. Ils proposent également une échelle d'incapacité partiellement ordonnée.

MOTS CLÉS: Echelle d'incapacité; analyse discriminante.

1. INTRODUCTION

Des efforts considérables ont été faits pour mieux connaître la population des personnes handicapées. Ces efforts se sont traduits surtout par l'élaboration de méthodes permettant de repérer cette population et par l'analyse de données d'enquête visant à faire mieux comprendre les divers aspects de l'incapacité et à mettre au point des mesures efficaces du degré d'incapacité. Citons, à titre d'exemple, l'ouvrage de Dolson *et coll.* (1984) et celui de Raymond *et coll.* (1981). Cet article décrit l'élaboration d'une méthode préliminaire visant à tracer un profil plus juste de la population souffrant d'incapacité au Canada. Nous nous livrons, en particulier, à une analyse typologique fondée sur les résultats de nombreuses analyses discriminantes.

La section suivante contient des renseignements sur l'Enquête sur la santé et l'incapacité au Canada. Dans la troisième section, nous décrivons la formation des grappes. La quatrième section porte sur la caractérisation de ces grappes tandis que la section 5 contient une analyse de leurs caractéristiques. Enfin, la section 6 sert de conclusion.

2. RENSEIGNEMENTS GÉNÉRAUX

Devant l'absence de données sur les personnes handicapées au Canada, Statistique Canada a mis en oeuvre un programme visant à créer une base de données sur l'incapacité. L'Enquête sur la santé et l'incapacité au Canada (ESIC) a été intégrée comme supplément à l'Enquête sur la population active du Canada (EPA) en octobre 1983 et juin 1984. Dans les deux cas, des questionnaires distincts avaient été prévus pour les enfants et pour les adultes. Dans l'enquête d'octobre, le questionnaire réservé aux adultes devait être rempli par toutes les unités de l'échantillon de l'EPA (la base de sondage comprend environ 97% de toute la population canadienne âgée de 15 ans ou plus). Dans l'enquête de juin, le questionnaire réservé aux adultes visait uniquement les personnes de 15 à 64 ans qui demeuraient dans une

¹ Il s'agit ici d'une version révisée de l'article qui a été présenté à la réunion de l'ASA, Social Statistics Section à Las Vegas en août 1985.
² D.A. Binder et G. Lazarus, Division des méthodes d'enquêtes sociales, Secteur de l'informatique et de la méthodologie, Statistique Canada, 4 étage, Immeuble Jean-Talon, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

- RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley.
- ROBINSON, P.M. (1982). On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics*, 24, 234-238.
- SMITH, T.M.F. (1981). Regression analysis for complex surveys. Dans *Current Topics in Survey Sampling*, (éd. D. Krewski, R. Platek, et J.N.K. Rao), New York: Academic Press, 267-292.
- THEIL, H. (1971). *Principles of Econometrics*. New York: Wiley.
- VAN PRAAG, B.M.S. (1981). Model-free regression. *Economics Letters*, 7, 139-144.
- VAN PRAAG, B.M.S. (1982). The population-sample decomposition with an application to minimum distance estimators. Rapport n° 8218, Centre de recherche en économie publique, Université Leyden.
- WHITE, H. (1980a). Nonlinear regression on cross section data. *Econometrica*, 48, 721-746.
- WHITE, H. (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, 12, 149-170.

Ten Cate: Analyse de régression pour des données d'enquête 142

Nous avons aussi considéré la régression sans modèle, laquelle exige la même pondération que la stratification endogène. Dans ce type de régression, la variance de l'estimateur des coefficients de régression n'est formée que de l'élément lié à l'échantillonnage, celui lié au modèle étant éliminé.

Enfin, nous avons présenté des considérations d'ordre pratique sur la pondération des données.

REMERCIEMENTS

L'auteur tient à remercier Abby Israëls, Albert Verbeek et plusieurs arbitres anonymes pour les commentaires qu'ils ont faits sur des versions antérieures de ce document.

BIBLIOGRAPHIE

- BETHLEHEM, J.G., et KELLER, W.J. (1983). Weighing sample survey data using linear models. Rapport interne, Département des méthodes statistiques, Bureau central de la statistique des Pays-Bas, Voorburg.
- BINDER, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BRESLOW, N., et DAY, N.E. (1980). *Statistical Methods in Cancer Research, Volume I: the Analysis of Case-Control Studies*. Centre international de recherche sur le cancer, Lyon.
- BREWER, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.
- CRAMER, J.S. (1971). *Empirical Econometrics*. Amsterdam: North-Holland.
- DUMOUCHEL, W.H., et DUNCAN, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, C37, 117-132.
- HAUSMAN, J.A., et WISE, D.A. (1981). Stratification on endogenous variables and estimation: the Gary income maintenance experiment. Dans *Structural Analysis of Discrete Data with Econometric Applications*, (éd. C.F. Manski et D. McFadden), Cambridge: MIT Press.
- HECKMAN, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-161.
- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- JEWELL, N.P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, 72, 11-21.
- JOHNSTON, J. (1972). *Econometric Methods*. Tokyo: McGraw-Hill Kogakusha.
- JONRUP, H., et RENNERMALM, B. (1976). Regression analysis in samples from finite populations. *Scandinavian Journal of Statistics*, 33-36.
- KMENTA, J. (1978). *Elements of Econometrics*. New York: McMillan.
- MANISKI, C.F., et McFADDEN, D. (éd.) (1981). *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: MIT Press.
- NATHAN, G. (1981). L'inférence statistique basée sur des plans d'échantillonnage complexes. *Techniques d'enquête*, 7, 109-130.
- PORTER, R.D. (1973). On the use of survey sample weights in the linear model. *Annals of Economic and Social Measurement*, 2, 141-158.

Le dernier membre de l'équation (68) correspond à la p -composante de la variance définie par l'équation (37). Nous pouvons en conclure que, dans la régression sans modèle, la variance de l'estimateur des coefficients de régression est constituée uniquement de la p -composante, la ξ -composante étant éliminée.

Enfin, sur la base des commentaires exprimés à la fin de la section 5, signalons que le dernier membre de l'équation (68) peut être exprimé comme suit,

$$(69) \quad (X'X)^{-1} \Sigma (X'X)^{-1},$$

où la matrice Σ est la p -variance-covariance des totaux de ligne de X' diag(e). Binder (1983, section 4) en arrive à la même conclusion mais selon un raisonnement différent.

8. ANALYSE

Cette section présente des considérations d'ordre pratique sur l'utilisation de poids dans l'analyse de régression. Nous allons examiner brièvement plusieurs raisons qui justifient l'utilisation de poids dans ce genre d'analyse, en regard à ce qui a été traité dans les sections précédentes.

En premier lieu, signalons qu'il peut y avoir une grande différence entre une régression pondérée et une régression non pondérée. Cela est particulièrement vrai lorsque les unités étudiées sont des entreprises commerciales, soit des exploitations agricoles, des entreprises industrielles ou tout autre genre d'entreprises commerciales très différentes au point de vue du nombre d'employés. Au Bureau central de la statistique des Pays-Bas, par exemple, le nombre d'employés est un critère de stratification standard dans les plans d'échantillonnage appliqués aux entreprises commerciales, ce qui crée une très grande diversité de probabilités d'inclusion, les grosses unités ayant relativement plus de chances d'être choisies. Dans les études où la variable endogène est l'emploi, le plan de sondage est nécessairement endogène et, de ce fait, exige une régression pondérée, par laquelle les grosses unités ont des poids peu élevés.

En deuxième lieu, lorsque l'analyse de régression porte sur des unités qui diffèrent beaucoup par leur taille, l'hétéroscédasticité des résidus pose un problème majeur. Il faut alors recourir à la régression pondérée, de même que nous l'avons fait dans la section 2 à cause de l'existence d'un plan endogène. On attribue alors des poids peu élevés aux grosses unités.

Enfin, la troisième raison qui justifie la pondération des données d'échantillon concerne la régression sans modèle, qui a été analysée dans la section 7. Dans ce cas, les poids sont du même genre que ceux définis dans la section 2.

En définitive, il ne semble y avoir aucune raison qui empêche d'incorporer le plan de sondage à l'analyse de régression.

9. CONCLUSION

Cette étude nous a permis d'analyser l'estimation d'un modèle de régression au moyen de données d'une enquête par sondage. Nous avons étudié, plus particulièrement, les échantillons prélevés à l'aide d'un plan de sondage endogène, par exemple des échantillons stratifiés en fonction de la variable endogène. Nous avons vu dans ce cas que la pondération des observations de l'enquête par l'inverse de la racine carrée des fractions de sondage produisait un estimateur convergent. La notion de convergence appliquée ici est une version modifiée de la notion proposée par Brewer (1979). Nous avons défini la variance asymptotique de l'estimateur de même qu'un estimateur convergent de cette variance, celle-ci étant la somme d'un élément lié à l'échantillonnage et d'un élément lié au modèle.

distribution de probabilité quelconque. Voir aussi DuMouchel et Duncan (1983). White (1980b, section 3) analyse des sujets connexes. En deuxième lieu, disons que l'estimateur par régression d'un agrégat de la population repose sur la régression sans modèle. Consulter des ouvrages comme celui de Cochran (1977), les comptes rendus de Nathan (1981) et de Smith (1981) et l'ouvrage de Bethlehem et Keller (1983).

La régression sans modèle sert à estimer le vecteur des paramètres d'une population

$$b = (X'X)^{-1}X'y, \quad (61)$$

en l'absence d'hypothèses concernant la distribution de probabilité de y . En fait, X et y sont considérées l'une et l'autre comme des variables non aléatoires. En outre, on utilise le même mode d'itération que dans la section 2, c'est-à-dire,

$$X = \iota_K \otimes X_0, \quad (62)$$

$$y = \iota_K \otimes y_0, \quad (63)$$

et

où y_0 désigne un vecteur fixe de dimension N_0 . Comme précédemment, les K matrices diagonales T_k ($k = 1, \dots, K$) sont indépendantes et identiquement distribuées. Comme dans la section 2, elles décrivent l'échantillon. Les matrices T_k constituent la matrice T . Aucune autre hypothèse n'est faite sur la distribution de T .

Nous pouvons démontrer assez facilement, un peu comme nous l'avons fait dans la section 2, que l'estimateur pondéré $\hat{\beta}$ défini en (7) est un estimateur convergent de b défini par l'équation (61). Voir aussi Jönrup et Rennermalin (1976), qui identifient $\hat{\beta}$ comme un estimateur "approximativement sans biais" de b , et Van Praag (1982, Section 4d), qui analyse, dans le contexte de la régression sans modèle, le "biais dû à la sélection" lorsque les probabilités d'inclusion sont connues.

Comme dans la section 4, il s'ensuit que la variance asymptotique de $\hat{\beta}$, désignée par $\text{Var}^{\text{SM}}(\hat{\beta})$, est définie de la façon suivante pour la régression sans modèle.

$$\text{Var}^{\text{SM}}(\hat{\beta}) = (X'X)^{-1}X'VX(X'X)^{-1}, \quad (64)$$

Notons que

$$e \equiv y - Xb, \quad (65)$$

$$V = \text{diag}(e) \Pi^{-1} P \Pi^{-1} \text{diag}(e), \quad (66)$$

et que P est défini comme dans l'équation (29). De plus, l'équation (66) diffère de l'équation (28) par l'élimination de la ξ -espérance et la substitution de e à ϵ .

Il serait intéressant de reformuler $\text{Var}^{\text{SM}}(\hat{\beta})$ comme nous l'avons fait pour $\text{Var}(\hat{\beta})$ dans la section 5. Pour cela, nous utiliserons

$$X'e = 0, \quad (67)$$

qui découle directement des équations (61) et (65). Ainsi, $\text{Var}^{\text{SM}}(\hat{\beta})$ peut être reformulée comme suit

$$\begin{aligned} \text{Var}^{\text{SM}}(\hat{\beta}) &= (X'X)^{-1}X' \text{diag}(e) (\Pi^{-1} P \Pi^{-1} - \iota \iota') \text{diag}(e) X(X'X)^{-1} \\ &\quad + (X'X)^{-1}X'e e' X(X'X)^{-1} \\ &= (X'X)^{-1}X' \text{diag}(e) (\Pi^{-1} P \Pi^{-1} - \iota \iota') \text{diag}(e) X(X'X)^{-1}. \end{aligned} \quad (68)$$

Dans l'échantillonnage stratifié, la probabilité d'inclusion de n importe quelle paire d'éléments d'une population qui n appartiennent pas à la même strate (probabilité d'inclusion du second ordre) est égale au produit des probabilités d'inclusion respectives de ces éléments (probabilité d'inclusion du premier ordre): l'inclusion du premier élément et l'inclusion du deuxième élément sont des événements indépendants. Cette règle s'applique à peu près identiquement à n importe quelle paire d'éléments d'une population qui appartiennent à la même strate. Ainsi, les éléments non diagonaux de P sont approximativement égaux aux éléments non diagonaux de $\Pi \Pi'$. Comme précédemment, P et Π ont la même diagonale. Donc, d'une manière approximative,

$$P = \Pi \Pi' / \Pi - \Pi^2 + \Pi. \tag{56}$$

Alors

$$V = E_{\xi} [\text{diag}(\epsilon)(\epsilon)' - I + \Pi - \Pi'] \text{diag}(\epsilon)]$$

$$= E_{\xi} [\epsilon \epsilon' - \text{diag}^2(\epsilon) + \text{diag}^2(\epsilon) \Pi - \Pi^{-1}] = E_{\xi} [\text{diag}^2(\epsilon) \Pi - \Pi^{-1}], \tag{57}$$

compte tenu de l'hypothèse (4). Dans le cas présent, V est donc une matrice diagonale. Alors

$$T \left(\frac{P}{V} \right) T' = T \Pi^{-1} E_{\xi} [\text{diag}^2(\epsilon) \Pi - \Pi^{-1}], \tag{58}$$

qui est aussi une matrice diagonale. Considérons maintenant un élément i de la population, qui appartient à l'échantillon. Alors, en appliquant l'équation (58) et en supposant une distribution normale des perturbations,

$$\left[T \left(\frac{P}{V} \right) T' \right]_{ii} = \frac{1}{H} \sum_{h=1}^H \frac{\pi_h}{1} \int_{L_h - x_i' \beta}^{L_h - 1 - x_i' \beta} \phi(\epsilon_i; \delta_2^i) \epsilon_i^2 d\epsilon_i$$

$$\delta_2^i \frac{\pi_{ii}}{1} + \frac{\pi^{(H)}}{1} \left\{ \sum_{h=1}^H \frac{\pi^{(h)}}{1} - \frac{\pi^{(h+1)}}{1} \right) \Phi [(L_h - x_i' \beta) / \hat{\sigma}] \Bigg\}. \tag{59}$$

Dans l'équation précédente, $\phi(\cdot; \delta_2^i)$ désigne la fonction de distribution d'une Loi normale, de moyenne nulle et de variance $\hat{\sigma}^2$. La fonction $\Phi(\cdot)$ est définie

$$\Phi(x) \equiv \int_x^{-\infty} \phi(\epsilon; 1) \epsilon^2 d\epsilon = \Phi(x) - x \phi(x; 1), \tag{60}$$

où $\Phi(\cdot)$ représente la fonction de distribution d'une Loi normale centrée réduite. Pour en arriver à l'équation (59), nous avons utilisé $\Phi(L_0) = 0$ et $\Phi(L_H) = 1$.

7. RÉGRESSION SANS MODÈLE

7.1 Estimateur convergent

Dans cette section, nous nous éloignons du sujet principal de notre analyse pour traiter de la régression sans modèle. Tout d'abord, disons que la régression sans modèle peut être utile lorsqu'on doute de la validité d'un modèle linéaire. Se référer à Fuller (1975), qui étudie la régression sans modèle pour des plans de sondage particuliers. Van Praag (1981, 1982) examine la régression sans modèle dans le cas d'un échantillonnage répété appliqué à une

Dans l'équation ci-dessus, P_0 désigne $E_p(T_k)$, qui est la même pour tous les $k = 1, \dots, K$. Le dernier signe d'égalité résulte de l'application du théorème de Khintchine, puisque les termes de la seconde sommation sur k sont indépendants et différemment distribués avec une p -espérance égale à $X_0' V_0 X_0$. Enfin, en combinant les équations (50), (51) et (53), nous obtenons

$$\text{plim } K \text{var}(\beta) = (X_0' X_0)^{-1} X_0' V_0 X_0 (X_0' X_0)^{-1}, \quad (54)$$

ce qui correspond au membre de droite de l'équation (49).

6.2 Échantillonnage stratifié

Ici, le calcul de la matrice $T(V/P)T$ est fait pour un cas particulier: 1) les perturbations sont distribuées suivant une loi normale et 2) le plan d'échantillonnage est fondé sur une stratification endogène de telle manière que la probabilité d'inclusion π_{ii} de l'élément i de la population est uniquement fonction du i -ème élément de y , par exemple de $y_{(i)}$. Ainsi,

$$\pi_{ii} = f(y_{(i)}), \quad (55)$$

pour $i = 1, \dots, N$. À titre d'exemple, prenons l'échantillon stratifié qui est illustré à la figure 1. Le plan d'échantillonnage, en l'occurrence, comprend trois strates. Les éléments de la strate médiane sont ceux pour lesquels la probabilité d'inclusion est la plus élevée. La figure 2 illustre la fonction f correspondante.

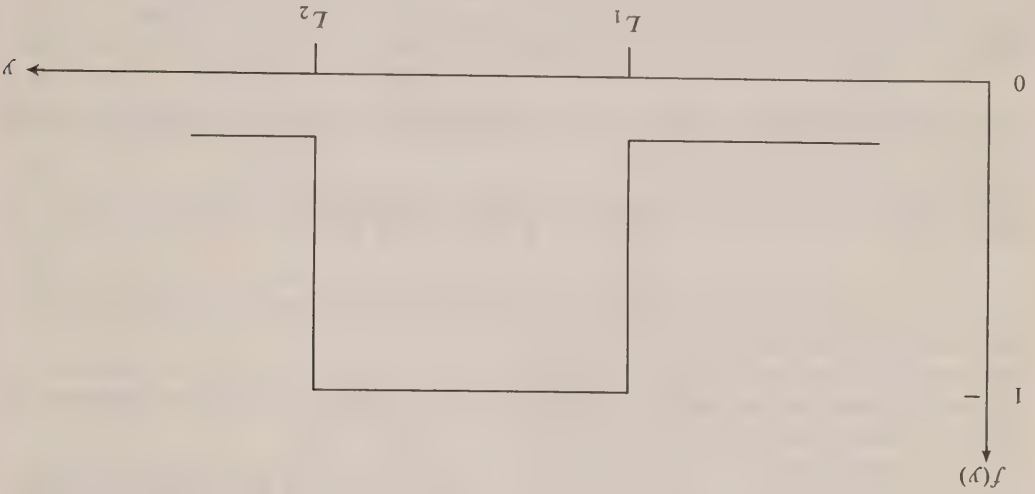


Figure 2. Fonction de probabilité f correspondant à la figure 1.

D'une manière générale, supposons H strates, désignées respectivement $h = 1, \dots, H$. Définissons L_0, L_1, \dots, L_H comme les bornes de ces strates. Normalement, $L_0 = -\infty$ et $L_H = +\infty$. Posons $\pi_{(h)}$ comme la probabilité d'inclusion des éléments de la population dans la strate h . De façon plus formelle, la fonction $f(\cdot)$ est telle que $f(y)$ égale $\pi_{(h)}$ si $L_{h-1} \leq y < L_h$. Dans la pratique, on connaît habituellement les valeurs de $\pi_{(h)}$ et de L_h puisque le plan de sondage utilisé en dépend.

où (V/P) désigne la matrice formée des quotients des éléments de V par les éléments correspondants de P .
Démonstration: Considérons tout d'abord la structure de V . Partitionnons la matrice V en une matrice carrée $K \times K$ formée de blocs de dimension $(N_0 \times N_0)$. Le (k, r) -ième bloc non diagonal de V est égal à

$$E_{\xi}[\text{diag}(\epsilon_k) \Pi_k^{-1} E_p(T_k \iota' T_r) \Pi_r^{-1} \text{diag}(\epsilon_r)] \\ = E_{\xi}[\text{diag}(\epsilon_k) \Pi_k^{-1} E_p(T_k) \iota' E_p(T_r) \Pi_r^{-1} \text{diag}(\epsilon_r)] \\ = E_{\xi}(\epsilon_k \epsilon_r') = 0, \tag{47}$$

si l'on applique le mode d'itération de la population et le plan d'échantillonnage proposés antérieurement. Les blocs diagonaux de V sont identiques et dépendent de X_0 . Par conséquent, $V(\beta, \sigma^2; X)$ peut s'exprimer comme suit:

$$V(\beta, \sigma^2; X) = I_K \otimes V_0(\beta, \sigma^2; X_0), \tag{48}$$

où $V_0(\beta, \sigma^2; X_0)$ est une fonction matricielle $N_0 \times N_0$. En nous servant des équations (1) et (48), nous pouvons reformuler $K\text{Var}(\beta)$ de la façon suivante,

$$K\text{Var}(\beta) = (X_0' X_0)^{-1} X_0' V_0 X_0 (X_0' X_0)^{-1}, \tag{49}$$

où V_0 désigne $V_0(\beta, \sigma^2; X_0)$. Le membre de droite de l'équation (49) est indépendant de K et par conséquent, égal à sa limite lorsque K tend vers l'infini. Considérons maintenant le membre de gauche de l'équation (45).

$$K\text{var}(\beta) = \left(\frac{1}{K} X' \Pi^{-1} T X \right)^{-1} \left[\frac{1}{K} X' T \left(\frac{P}{P'} \right) T X \right] \left(\frac{1}{K} X' \Pi^{-1} T X \right)^{-1}. \tag{50}$$

Pour déterminer les équations (13) et (22), nous avons utilisé l'équation

$$\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T X \right) = X_0' X_0. \tag{51}$$

Donc, compte tenu de l'hypothèse voulant que $V(\beta, \sigma^2; X)$ soit une fonction continue, nous avons

$$\text{plim}_{K \rightarrow \infty} V_0 = V_0, \tag{52}$$

où V_0 désigne $V_0(\beta, \sigma^2; X_0)$. Par les équations (1), (48) et (52), nous avons

$$\text{plim}_{K \rightarrow \infty} \frac{1}{K} X' T \left(\frac{P}{P'} \right) T X = \text{plim}_{K \rightarrow \infty} \frac{1}{K} \sum^k \left[X_0' T_k \left(\frac{P_0}{P_0'} \right) T_k X_0 \right] \\ = \text{plim}_{K \rightarrow \infty} \frac{1}{K} \sum^k \left[X_0' T_k \left(\frac{P_0}{P_0'} \right) T_k X_0 \right] = X_0' V_0 X_0. \tag{53}$$

Étant donné l'équation (39), il s'agit évidemment d'un estimateur sans biais. La variance de X est

$$\text{Var}(X) = E_p(X^2) - [E_p(X)]^2 = E_p(x' \Pi^{-1} T \Pi^{-1} T \Pi^{-1} x) - x' \Pi^{-1} x$$

$$= x' (\Pi^{-1} P \Pi^{-1} - \Pi^{-1}) x. \quad (42)$$

Le dernier membre de l'équation (42) est la formule de la variance de l'estimateur d'Horvitz-Thompson, que l'on retrouve dans des ouvrages sur l'échantillonnage, comme celui de Cochran (1977). La forme matricielle est toutefois peu usitée dans ces ouvrages. L'expression entre parenthèses dans le dernier membre de l'équation (38), qui définit V^* , est contenu dans la formule de la p -composante de $\text{Var}(\beta)$, les éléments diagonaux de la p -composante de la matrice de variances $\text{Var}(\beta)$ peuvent être considérés comme la ξ -espérance de la p -variance de l'estimateur d'Horvitz-Thompson des totaux de ligne de $(X'X)^{-1} X' \text{diag}(\epsilon)$. Ces totaux sont les éléments du vecteur $(X'X)^{-1} X' \epsilon$.

6. ESTIMATION DE $\text{VAR}(\beta)$

6.1 Cas général

Cette section porte sur l'estimateur de la variance asymptotique $\text{Var}(\beta)$. Il est plutôt difficile de trouver un estimateur convergent de $\text{Var}(\beta)$ puisque, pour cela, il faut connaître la relation F qui existe entre y et le plan de sondage, telle qu'elle est définie dans la matrice V . En pratique, seul le plan d'échantillonnage pour les valeurs réelles de y peut être connu. En règle générale, le plan ne permet pas, à lui seul, de prévoir ce qu'il deviendrait si y prenait d'autres valeurs. Dans une certaine mesure, on ne parle pas uniquement d'un modèle de régression mais aussi d'un modèle du concept même du plan!

Pour le moment, nous supposons que la fonction F est connue et, par conséquent, V est une fonction connue de X et des paramètres du modèle. (Voir la section 6.2 pour un cas particulier.) La fonction V s'exprime comme suit:

$$V = V(\beta, \sigma^2; X),$$

(43)

Nous supposons que $V(\beta, \sigma^2, X)$ est une fonction continue. Par souci de concision, nous définissons V

$$V \equiv V(\beta, \sigma^2; X),$$

(44)

où β et σ^2 sont des estimateurs convergents de β et de σ^2 respectivement. Le reste de cette section est consacré à la démonstration d'un théorème concernant un estimateur convergent de la variance.

Dans le cas présent, un estimateur $\text{Var}(\beta)$ est dit convergent si:

$$\text{plim } K \text{Var}(\beta) = \lim_{K \rightarrow \infty} K \text{Var}(\beta).$$

(45)

Théorème 4. Suivant les hypothèses énoncées ci-dessus, un estimateur convergent de la variance asymptotique $\text{Var}(\beta)$ est donné par

$$\text{var}(\hat{\beta}) = (X' \Pi^{-1} T X)^{-1} X' T \left(\frac{D}{d} \right) T X (X' \Pi^{-1} T X)^{-1}, \quad (46)$$

$$\text{Var}(\delta) = \frac{1}{k} \sum E_p^{\delta} E_p^{\delta'}$$

$$= \frac{1}{k} (X_0' X_0)^{-1} \left[E_p^{\delta} E_p^{\delta'} \right] \sum_{k=0}^k X_0' \Pi_k^{-1} T_{k \epsilon k} T_{k \Pi k}^{-1} X_0 \left(X_0' X_0 \right)^{-1}$$

$$= K(X'X)^{-1} [E_p^{\delta} E_p^{\delta'} (X' \Pi^{-1} T \epsilon T' \Pi^{-1} X)] (X'X)^{-1}$$

$$= K(X'X)^{-1} X' \{ E_p^{\delta} E_p^{\delta'} \} \text{diag}(\epsilon) \Pi^{-1} T \Pi^{-1} T' \Pi^{-1} \text{diag}(\epsilon)] X(X'X)^{-1}$$

$$= K(X'X)^{-1} X' \{ E_p^{\delta} \} \text{diag}(\epsilon) \Pi^{-1} E_p(T \Pi^{-1} T') \Pi^{-1} \text{diag}(\epsilon)] X(X'X)^{-1} \quad (36)$$

En divisant $\text{Var}(\delta)$ par K , on obtient $\text{Var}(\delta)$ ce qui conclut la démonstration.

5. DÉCOMPOSITION DE VAR(δ)

L'équation (27) peut être formulée autrement:

$$\text{Var}(\delta) = \sigma^2 (X'X)^{-1} + (X'X)^{-1} X' V^* X (X'X)^{-1} \quad (37)$$

où, par l'équation (4),

$$V^* \equiv E_{\xi} [\text{diag}(\epsilon) (\Pi^{-1} P \Pi^{-1} - \Pi^{-1} \text{diag}(\epsilon))], \quad (38)$$

Le premier terme du membre de droite de l'équation (37) pourrait être désigné à bon droit comme la ξ -composante de la variance de $\hat{\beta}$. Cette composante représenterait, à elle seule, la variance de $\hat{\beta}$ si toute la population était échantillonnée. Elle dépend entièrement des variations de la perturbation ϵ et est l'expression habituellement utilisée pour ce genre de situation. Le deuxième terme du membre de droite de l'équation (37) peut être désigné la p -composante de la variance de $\hat{\beta}$. Cette composante renferme les matrices Π et P , qui décrivent le plan de sondage. Elle ressemble à la formule de la variance de l'estimateur d'un agrégat ou d'une moyenne d'une population finie. Pour mieux comprendre la signification de la p -composante de $\text{Var}(\hat{\beta})$, nous analysons brièvement dans cette section la théorie sur laquelle repose ce genre d'estimateur.

Nous considérons une population finie de N éléments. (Dans le cas présent, on ne suppose pas l'existence d'un mode d'itération.) À chaque élément de la population correspond une valeur d'une variable non aléatoire réelle, qui est tirée d'un N -vecteur x . Un échantillon est prélevé, sans remise, dans cette population. Comme précédemment, l'échantillon est défini par la matrice diagonale T . De même,

$$\Pi \equiv E_p(T) \quad (39)$$

$$P \equiv E_p(T \Pi^{-1} T'), \quad (40)$$

ces deux équations désignant respectivement les probabilités d'inclusion du premier ordre et les probabilités d'inclusion du second ordre. Comme, dans ce cas-ci, il n'y a pas de modèle de régression, Π et P sont des matrices dont les éléments sont connus et fixes. Horvitz et Thompson (1952) ont proposé d'estimer l'agrégat de population $X'x$ par

$$X' = x' \Pi^{-1} T' \quad (41)$$

et

$$P \equiv E_p(Tu'T) \tag{29}$$

Les éléments de P sont désignés les probabilités d'inclusion du second ordre, c'est-à-dire la probabilité qu'une paire quelconque d'éléments de la population soit incluse dans l'échantillon. P a la même diagonale que Π . Le reste de la section est consacré à la démonstration du théorème.

Démonstration: Considérons la distribution asymptotique (lorsque $K \rightarrow \infty$) de

$$K_{1/2}(\hat{\beta} - \beta) = K_{1/2}[(X' \Pi^{-1} T X)^{-1} X' \Pi^{-1} T y - \beta]$$

$$= K_{1/2}(X' \Pi^{-1} T X)^{-1} X' \Pi^{-1} T \epsilon \tag{30}$$

Puisque

$$\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T X \right) = X_0' X_0 \tag{31}$$

$K_{1/2}(\hat{\beta} - \beta)$ a la même distribution asymptotique que δ , ce dernier étant défini

$$\delta \equiv K^{-1/2}(X_0' X_0)^{-1} X_0' \Pi^{-1} T \epsilon = K^{-1/2} \sum_{k=1}^K X_0' \Pi^{-1} T_k \epsilon_k = K^{-1/2} \sum_{k=1}^K \delta_k \tag{32}$$

et

$$\delta_k \equiv (X_0' X_0)^{-1} X_0' \Pi^{-1} T_k \epsilon_k \tag{33}$$

pour tous les $k = 1, \dots, K$. (Voir, par exemple, Rao (1973), p. 122.) Comme les vecteurs δ_k ($k = 1, \dots, K$) sont indépendants et identiquement distribués et qu'en plus

$$E_{\xi} E_p(\delta_k) = (X_0' X_0)^{-1} X_0' E_{\xi} E_p(\Pi^{-1} T_k \epsilon_k)$$

$$= (X_0' X_0)^{-1} X_0' E_{\xi} [\Pi^{-1} E_p(T_k) \epsilon_k]$$

$$= (X_0' X_0)^{-1} X_0' E_{\xi}(\epsilon_k) = 0, \tag{34}$$

la variance de δ , désignée $\text{Var}(\delta)$, est la même pour tous les K et est égale à la variance de la distribution asymptotique de δ lorsque $K \rightarrow \infty$. Elle peut être exprimée comme suit

$$\text{Var}(\delta) = E_{\xi} E_p(\delta_k \delta_k') \tag{35}$$

pour tout $k \in \{1, \dots, K\}$. Puisque les vecteurs δ_k sont indépendants et identiquement distribués, l'équation ci-dessus peut être reformulée comme suit

$$\text{plim}_{K \rightarrow \infty} \left[\frac{1}{N} K' X (X' X)^{-1} X' \varepsilon \right]$$

$$= \left[\text{plim}_{K \rightarrow \infty} \left(\frac{1}{N} K' \varepsilon \right) \right]' \left[\text{plim}_{K \rightarrow \infty} \left(\frac{1}{N} X' X \right) \right]^{-1} \text{plim}_{K \rightarrow \infty} \left(\frac{1}{N} K' X' \varepsilon \right)$$

$$= \left[\text{plim}_{K \rightarrow \infty} \left(\frac{1}{N} X' \Pi^{-1} T \varepsilon \right) \right]' \left[\text{plim}_{K \rightarrow \infty} \left(\frac{1}{N} X' \Pi^{-1} T X \right) \right]^{-1} \text{plim}_{K \rightarrow \infty} \left(\frac{1}{N} K' X' \Pi^{-1} T \varepsilon \right)$$

$$(22) \qquad = 0' (X_0' X_0)^{-1} 0 = 0.$$

Pour en arriver à l'équation (22), nous avons appliqué le lemme 1 de la même manière que dans la détermination de l'équation (13). En combinant les équations (20), (21) et (22), nous obtenons

$$(23) \qquad \text{plim}_{K \rightarrow \infty} \left(\frac{1}{N} K' \varepsilon \right) = N_0 \sigma^2.$$

Enfin, nous appliquons le lemme 1 au premier facteur de l'équation (14) en remplaçant z et η , dans la formule du lemme, par ι_N . Nous obtenons donc

$$(24) \qquad \text{plim}_{K \rightarrow \infty} \left(\frac{1}{N} \iota_N' \Pi^{-1} T \iota_N \right) = N_0.$$

Etant donné les équations (23) et (24), nous avons

$$(25) \qquad \text{plim}_{K \rightarrow \infty} (\sigma^2) = \sigma^2,$$

ce qui démontre le théorème. Enfin, il peut être utile de signaler, comme corollaire de l'équation (23), que

$$(26) \qquad \left(\frac{1}{N} \right) \varepsilon' \varepsilon$$

est un estimateur convergent de σ^2 .

4. VARIANCE DE $\hat{\beta}$

Dans cette section, nous définissons la variance asymptotique de l'estimateur $\hat{\beta}$.
Théorème 3. La variance asymptotique de $\hat{\beta}$ est définie

$$(27) \qquad \text{Var}(\hat{\beta}) = (X' X)^{-1} X' V X (X' X)^{-1},$$

où

$$(28) \qquad V \equiv E_{\varepsilon} [\text{diag}(\varepsilon) \Pi^{-1} P \Pi^{-1} \text{diag}(\varepsilon)],$$

3. ESTIMATION DE LA VARIANCE DE LA PERTURBATION

Le modèle de régression défini dans la section précédente contient deux paramètres: β et σ^2 . Le théorème 1 concernait l'estimation de β ; cette fois, nous nous penchons sur l'estimation de σ^2 . Dans la présente section, on cherche à démontrer le théorème suivant:

Théorème 2. La variance de l'échantillon pondéré des résidus de y est un estimateur convergent de la variance σ^2 de la perturbation si les poids de l'échantillon égalent l'inverse de la racine carrée des probabilités d'inclusion.

Démonstration: L'estimateur de la variance défini dans le théorème est

(14)
$$\hat{\sigma}^2 = (\iota_N \Pi^{-1} T \iota_N)^{-1} \hat{e}' \hat{e}$$

où

(15)
$$\hat{e} \equiv \Pi^{-1/2} T(y - X\beta).$$

Posons

(16)
$$y \equiv \Pi^{-1/2} Ty,$$

(17)
$$X \equiv \Pi^{-1/2} TX,$$

et

(18)
$$\hat{e} \equiv \Pi^{-1/2} T\hat{e}.$$

Alors

(19)
$$\hat{e} = y - X\hat{\beta} = y - X(\tilde{X}'\tilde{X})^{-1}\tilde{X}'y$$

et

$$\hat{e}'\hat{e} = y'[I_N - X(\tilde{X}'\tilde{X})^{-1}\tilde{X}']y = (X\beta + \varepsilon)'[I_N - X(\tilde{X}'\tilde{X})^{-1}\tilde{X}'](X\beta + \varepsilon)$$

$$= \varepsilon'\varepsilon - \varepsilon'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\varepsilon.$$

(20) La convergence en probabilité du premier terme du membre de droite de l'équation (20) s'exprime de la façon suivante,

$$\text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \varepsilon' \varepsilon \right) = \text{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \varepsilon' \Pi^{-1} T \varepsilon \right) = \text{plim}_{K \rightarrow \infty} \left[\frac{1}{K} \iota_N \Pi^{-1} T \text{diag}(\varepsilon) \varepsilon \right]$$

(21)
$$= \iota_{N_0}'(\sigma^2 \iota_{N_0}) = N_0 \sigma^2.$$

Dans cette équation, $\text{diag}(\varepsilon)$ désigne la matrice diagonale et les valeurs de ε constituent la diagonale principale. Pour en arriver à l'expression ci-dessus, nous nous sommes servis de l'équation (4) et avons appliqué la formule du lemme 1 en substituant ι_N à z et $\text{diag}(\varepsilon)$ à y . Considérons maintenant le second terme du membre de droite de l'équation (20).

Le reste de la section est consacré à la démonstration de ce théorème. Pour les besoins de cette démonstration, nous allons nous servir du lemme ci-dessous, qui sera aussi utilisé dans les démonstrations des théorèmes énoncés plus loin.

Lemme 1. Nous considérons un N -vecteur z , de telle sorte que $z = {}_{\iota_k} z_0$, où z_0 est un N_0 -vecteur fixe. Nous considérons également un N -vecteur η décomposé de telle sorte que $\eta' = (\eta^1, \eta^2, \dots, \eta^k)$. Chaque η^k possède N_0 éléments. Supposons que chaque η^k est une fonction de X_0, β et de ϵ^k , toutes ces fonctions étant identiques. Alors

$$(8) \qquad \operatorname{plim}_{K \rightarrow \infty} \left(\frac{1}{K} z' \Pi^{-1} T \eta \right) = z_0' E_{\xi}(\eta_0),$$

où $E_{\xi}(\eta_0)$ est l'espérance mathématique de n importe quelle valeur de η^k , la même pour tout k .

Démonstration du lemme 1: Considérons l'espérance de $\Pi^k {}_1 T^k \eta^k$:

$$(9) \qquad E_{\xi} E_p(\Pi^k {}_1 T^k \eta^k) = E_{\xi} [\Pi^k {}_1 E_p(T^k) \eta^k] = E_{\xi}(\eta^k),$$

pour tous les k . Comme la distribution de η^k est la même pour chaque k , nous pouvons écrire

$$(10) \qquad E_{\xi} E_p(\Pi^k {}_1 T^k \eta^k) = E_{\xi}(\eta_0)$$

pour tous les k . De plus, les K vecteurs $z_0' \Pi^k {}_1 T^k \eta^k$ sont indépendants et identiquement distribués. Par conséquent, le théorème de Khintchine s'applique comme suit,

$$\operatorname{plim}_{K \rightarrow \infty} \left(\frac{1}{K} z' \Pi^{-1} T \eta \right) = \operatorname{plim}_{K \rightarrow \infty} \left(\frac{1}{K} \sum^k z_0' \Pi^k {}_1 T^k \eta^k \right) = E_{\xi} E_p(z_0' \Pi^k {}_1 T^k \eta^k)$$

$$(11) \qquad = z_0' E_{\xi} E_p(\Pi^k {}_1 T^k \eta^k).$$

En appliquant l'équation (10) au membre de droite de l'équation (11), on obtient l'équation du lemme. Il est désormais facile de faire la démonstration du théorème 1.

Démonstration du théorème 1: L'estimateur des moindres carrés généralisé du théorème peut s'écrire de la façon suivante:

$$(12) \qquad \hat{\beta} = (X' \Pi^{-1} T X)^{-1} X' \Pi^{-1} T y = \beta + (X' \Pi^{-1} T X)^{-1} X' \Pi^{-1} T \epsilon.$$

Donc

$$\operatorname{plim}_{K \rightarrow \infty} \hat{\beta} = \beta + \left[\operatorname{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T X \right) \right]^{-1} \left[\operatorname{plim}_{K \rightarrow \infty} \left(\frac{1}{K} X' \Pi^{-1} T \epsilon \right) \right]$$

$$(13) \qquad = \beta + (X_0' X_0)^{-1} X_0' 0 = \beta.$$

L'expression $(X_0' X_0)$ est obtenue par l'application répétée du lemme 1, où les valeurs de z et de η sont remplacées par les colonnes de X . Il convient de souligner que $E_{\xi}(X_0) = X_0$ étant donné que X_0 est une constante. L'expression $X_0' 0$ est obtenue par l'application répétée du lemme 1, où les valeurs de z sont remplacées par les colonnes de X et celles de η par ϵ .

pour tout $i = 1, \dots, N$. Naturellement, T est idempotente. L'espace-échantillon S est l'ensemble de toutes les matrices T ainsi définies. Cet ensemble est défini. Le plan d'échantillonnage est une distribution de probabilité sur les éléments de l'espace-échantillon S . En l'occurrence, ce plan est endogène, c'est-à-dire qu'il dépend de y . Par conséquent, le plan de sondage proprement dit est stochastique. (Un plan qui ne dépend pas de y est appelé exogène ou non informatif.) Partitionnons T en une matrice carrée $K \times K$ formée de blocs $(N_0 \times N_0)$. Supposons que T_k désigne le k -ième bloc diagonal correspondant à la k -ième itération. De même, supposons que y est décomposé en K vecteurs de dimension N_0 , de telle manière que $y' = (y'_1, y'_2, \dots, y'_K)$. En outre, nous supposons que le plan de sondage tient à y dans le sens suivant: les K paires $(T_1, y_1), \dots, (T_K, y_K)$ sont indépendantes et identiquement distribuées. L'espérance de S , qui dépend de y (ou de ϵ), occupe une place importante dans cette étude. Cette espérance est désignée par E_p . Nous définissons donc

(5)
$$\Pi \equiv E_p(T).$$

Nous supposons que Π est connue. Ses éléments diagonaux sont appelés probabilités d'inclusion, c'est-à-dire la probabilité que les éléments d'une population soient inclus dans l'échantillon. La matrice Π est partitionnée en une matrice carrée $K \times K$ formée de blocs $(N_0 \times N_0)$. Définissons Π_k le k -ième bloc diagonal, correspondant à la k -ième itération. Souignons que chaque Π_k est stochastique puisqu'il dépend de y_k . Comme les K paires $(T_1, y_1), \dots, (T_K, y_K)$ sont indépendantes et identiquement distribuées, les blocs diagonaux Π_1, \dots, Π_K le sont aussi. La relation de dépendance qui existe entre Π_k et y est désignée par une fonction F , de sorte que

(6)
$$\Pi_k = F(y_k)$$

pour tous les $k = 1, \dots, K$. Nous supposons que $F(y_k)$ est non singulière pour chaque y_k . Autrement dit, les probabilités d'inclusion sont toujours positives. Ce modèle diffère quelque peu de celui de Brewer (1979). En effet, le modèle de Brewer ne comporte pas de variable endogène et, par conséquent, tous les blocs diagonaux Π_k sont égaux et non-aléatoires. On peut aussi comparer ce modèle avec la notion de "constante dans des échantillons itérés", que l'on trouve dans les ouvrages d'économétrie. Voir, par exemple, Theil (1971, p. 364). Nous sommes maintenant prêts à estimer β . Nous allons considérer les propriétés stochastiques des estimateurs pour toutes les paires $(y, T) \in (R^N_N \times S)$. L'espérance mathématique correspondante sera désignée par E_p . Nous considérons donc un estimateur des moindres carrés généralisé de β , disons $\hat{\beta}$, dont les poids sont égaux à la racine carrée des probabilités d'inclusion. Ainsi, nous avons

(7)
$$\hat{\beta} \equiv [(\Pi^{-1/2} X)' T (\Pi^{-1/2} X)]^{-1} (\Pi^{-1/2} X)' T (\Pi^{-1/2} y)$$

Il faut se rappeler que la matrice Π est connue. Souignons aussi que X et y ont trait à la population tandis que T opère une sommation sur les éléments de l'échantillon. Au lieu de considérer $\hat{\beta}$ comme un estimateur des moindres carrés généralisé, nous pouvons supposer que tous les éléments de π^{-1} sont des nombres entiers. Alors, si chaque observation i de l'échantillon est reproduite π^{-1}_i fois, $\hat{\beta}$ devient l'estimateur des moindres carrés ordinaire appliqué à cet échantillon grossi. Dans ce contexte, il n'est plus question de racine carrée des probabilités. Voir aussi Hausman et Wise (1981, p. 373). Le théorème principal de cette étude est le suivant.

Théorème 1. Suivant les hypothèses énoncées ci-dessus ((1), (2) et la distribution de ϵ et de T), l'estimateur des moindres carrés généralisé $\hat{\beta}$, défini dans l'équation (7), est convergent lorsque $K \rightarrow \infty$.

2. LE MODÈLE, L'ÉCHANTILLON ET UN ESTIMATEUR PAR RÉGRESSION

Dans cette section, nous étudions les propriétés asymptotiques d'un modèle de régression appliqué à l'échantillonnage sans remise d'une population finie. Il peut sembler contradictoire de parler de théorie asymptotique des échantillons prélevés sans remise dans une population finie puisque de tels échantillons doivent être bornés. La contradiction disparaît si l'on accroit dans la même proportion la taille de la population et celle de l'échantillon, celui-ci n'étant aucunement borné. L'interdépendance des probabilités d'inclusion des unités d'une population dans l'échantillon constitue un autre problème, surtout lorsqu'il s'agit de plans de sondage complexes. Pour résoudre ce problème, nous nous inspirons de Brewer (1979). Le modèle de Brewer nous permet d'utiliser des théorèmes de limites pour des suites de variables indépendantes puis d'appliquer néanmoins les résultats à des plans de sondage complexes. Essentiellement, ce modèle repose sur la notion d'itération évoquée précédemment. Cette notion sera très souvent appliquée dans cette étude. Pour une approche différente, voir Robinson (1982).

Tout d'abord, la structure de la population et le modèle sont connus. Considérons un ensemble fini de N_0 éléments. Chaque élément possède r caractéristiques non aléatoire exogènes de valeur réelle, formant pour l'ensemble fini d'éléments une matrice X_0 de dimension $(N_0 \times r)$. Nous posons l'une des hypothèses fondamentales de cette étude: la population est constituée de K itérations de l'ensemble fini de N_0 éléments, c'est-à-dire qu'elle possède $N \equiv KN_0$ éléments. La matrice des variables exogènes est X , où

(1)
$$X = \iota_K \otimes X_0.$$

Dans cette équation, ι_K est le K -vecteur dont tous les éléments égalent l'unité, et \otimes désigne le produit de Kronecker. On obtient des résultats asymptotiques en faisant tendre K vers l'infini. Les hypothèses du modèle décrivent un modèle linéaire standard. À chacun des N éléments de la population correspond une valeur de la variable aléatoire endogène. On obtient ainsi un N -vecteur y . Nous supposons que

(2)
$$E_{\xi}(y) = X\beta$$

pour un r -vecteur fixe et inconnu β . E_{ξ} désigne l'espérance mathématique pour toutes les valeurs $y \in R^N$. Définissons ensuite

(3)
$$\varepsilon = y - X\beta$$

Nous supposons que les N éléments de ε sont indépendants et identiquement distribués. D'après l'équation (2), nous pouvons affirmer que tous les éléments de ε ont une espérance mathématique nulle. Leur variance est σ^2 , c'est-à-dire

(4)
$$E_{\xi}(\varepsilon\varepsilon') = \sigma^2 I.$$

Dans le cas présent, l'échantillonnage est sans remise, conformément à la pratique courante. L'échantillon est défini par une matrice diagonale T de dimension $(N \times N)$ de telle manière que

$$t_{ii} = \begin{cases} 1 & \text{si l'élément } i \text{ de la population est inclus dans l'échantillon} \\ 0 & \text{dans le cas contraire;} \end{cases}$$

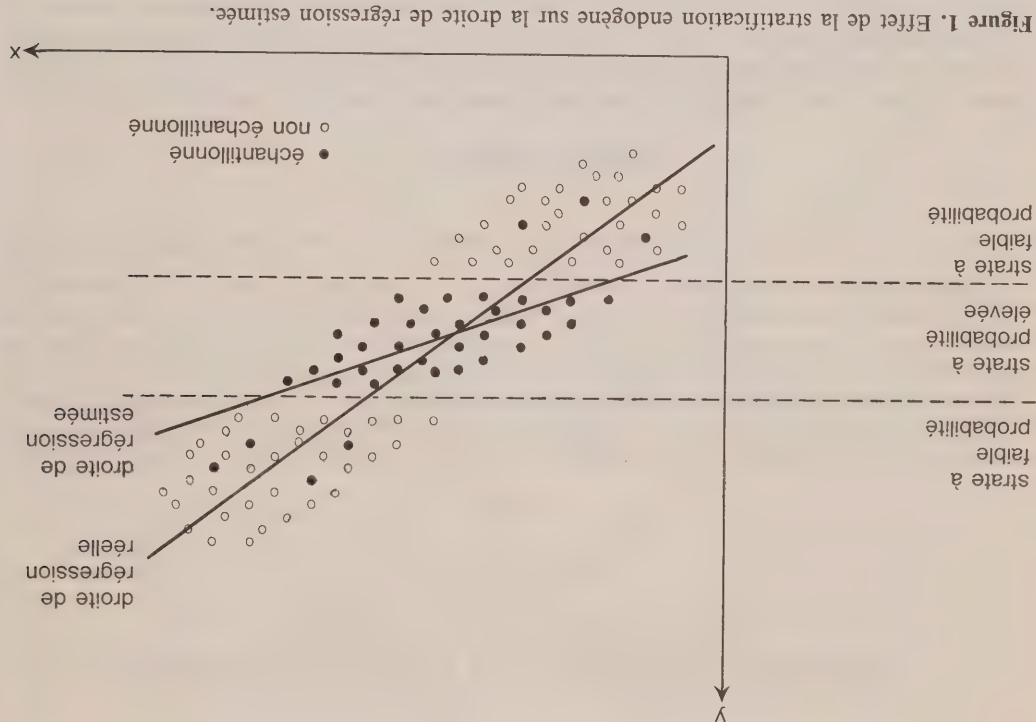


Figure 1. Effet de la stratification endogène sur la droite de régression estimée.

et Johnston (1972, section 9.2) examinent la question des variables explicatives aléatoires, qui est étroitement liée à la question précédente. Voir aussi White (1980a) pour la régression non linéaire. Le sujet qui nous occupe, soit l'analyse de la régression avec plan de sondage endogène, est plus complexe. Hausman et Wise (1981) étudient des plans de sondage stratifiés dans des conditions très simples: deux strates et un modèle de régression ne comportant qu'une constante. Jewell (1985) définit quelques estimateurs itératifs pour la stratification endogène.

L'analyse de régression avec plan de sondage endogène se rattache au problème de la non-réponse endogène dans l'analyse de régression normale. Voir Heckman (1979). Toutefois, l'analyse de régression avec plan endogène pose un peu moins de difficultés puisque les probabilités d'échantillonnage sont censées être connues; elles font partie intégrante du plan de sondage choisi. Par ailleurs, comme nous le verrons dans la section 6.1, il est plutôt difficile, en règle générale, d'estimer la variance lorsque nous avons un plan endogène.

L'analyse de régression avec plan de sondage endogène peut être comparée à l'analyse par la méthode des logits avec plan de sondage endogène, aussi appelée analyse par la méthode des logits avec échantillonnage fondé sur le choix ou échantillonnage de cas avec groupe témoin. Voir Manski et McFadden (1981, chapitres 1 et 2) et Breslow et Day (1980, section 6.3). La présente étude se divise comme suit. Dans les sections 2 et 3, nous énonçons les principaux théorèmes, lesquels définissent des estimateurs convergents des paramètres d'un modèle de régression linéaire; ces estimateurs sont établis à partir d'un échantillon prélevé au moyen d'un plan endogène. La convergence est définie à peu près de la même façon que dans l'analyse du biais faite plus haut, sauf que la définition est légèrement plus raffinée. On effectue plusieurs itérations des valeurs de x et on observe les valeurs de y obtenues selon le modèle de régression. Dans les sections 4 et 5, nous étudions la variance des estimateurs des coefficients de régression. Dans la section suivante, nous analysons l'estimation de ces variances. La section 7 porte sur la régression sans modèle tandis que la section 8 traite des divers motifs qui justifient l'utilisation de la régression pondérée. Enfin, la section 9 présente la conclusion.

Analyse de régression pour des données d'enquête avec plan de sondage endogène

ARIE TEN CATE¹

RÉSUMÉ

Cette étude traite de l'influence du plan de sondage sur l'estimation d'un modèle de régression linéaire. L'auteur étudie plus particulièrement les plans de sondage qui dépendent des valeurs de la variable endogène dans la population, c'est-à-dire les plans endogènes (ou informatifs). L'auteur définit aussi un estimateur convergent des coefficients de régression. La variance de cet estimateur correspond à la somme d'un élément lié au plan de sondage et d'un autre lié à la perturbation. L'auteur examine aussi brièvement la régression sans modèle. L'estimateur utilisé dans une régression de ce genre équivaut à l'estimateur utilisé dans une régression avec modèle lorsqu'il y a un plan de sondage endogène.

MOTS CLÉS: Régression; enquête par sondage; plan de sondage endogène.

1. INTRODUCTION

Tout modèle statistique repose fondamentalement sur l'hypothèse selon laquelle la valeur d'une ou de plusieurs variables est tirée d'une distribution de probabilité (par exemple, un modèle de régression avec des perturbations distribuées suivant une loi normale). Dans ce document, nous étudions un ensemble fini d'éléments qui répondent aux règles d'un modèle de ce genre. Cet ensemble d'éléments est appelé la population. Un échantillon est prélevé dans cette population, sans remise. Le présent document a pour objet d'étudier l'influence du plan de sondage sur l'estimation des paramètres du modèle. Cette influence variera essentiellement selon que le plan de sondage sera exogène ou endogène par rapport au modèle. Si le plan de sondage est endogène (ou informatif), les probabilités d'échantillonnage dépendront de la valeur des variables endogènes (dépendantes). Dans ce cas, le plan de sondage devra être pris en considération dans l'estimation des paramètres du modèle. La nature du problème est illustrée à la figure 1 où l'on peut voir un plan d'échantillonnage stratifié. Trois strates sont définies pour la variable endogène du modèle de régression. La strate médiane a une fraction de sondage plus élevée que les deux autres strates. Nous constatons que la pente de la droite de régression estimée, fondée uniquement sur des données échantillonnées, est biaisée par défaut si l'on fait abstraction du plan de sondage. Ce biais persiste avec de grands échantillons. Cela peut être vérifié intuitivement en imaginant que chacun des points (noirs ou blancs) de la figure 1 représente une multitude de points identiques. Même si cette multitude tend vers l'infini, la pente de la droite de régression estimée demeurera biaisée par défaut parce que la forme du nuage ne changera pas.

L'application des méthodes de régression à l'échantillonnage de populations finies fait l'objet d'études de plus en plus nombreuses qui portent sur des problèmes variés. Par exemple, on cherche à savoir comment appliquer des méthodes de régression à l'estimation de tels que $\Sigma xy / \Sigma x^2$, où la sommation porte sur tous les éléments de la population finie. Nathan (1981) et Smith (1981) passent en revue la littérature sur ce sujet. Un troisième problème auquel on s'intéresse est l'estimation des paramètres d'un modèle de régression au moyen d'un échantillon prélevé dans une population finie. Ce problème peut être résolu assez facilement lorsque le plan de sondage est exogène. Voir Porter (1973, section 1.2), DuMouchel et Duncan (1983) et des ouvrages comme celui de Cramer (1971, p.143). Kmenta (1971, section 8.3)

¹ Bureau central de planification, Van Stolkweg 14, 2585 JR La Haye, Pays-Bas.

Par contre, l'indice arithmétique et dans une certaine mesure l'indice interpolé ont tendance à être trop lisses. En d'autres termes, ils ont tendance à lisser tous les sommets en plus de faire ressortir un ou deux décalages. Bien que les indices marginal et géométrique ne soient pas entièrement à l'abri de ces décalages, ils ont tendance à suivre l'indice vrai d'un peu plus près. L'indice marginal donne le meilleur résultat d'ensemble, cependant, en raison de l'attrait mathématique de l'indice géométrique, c'est-à-dire l'indépendance théorique de son passé et sa correspondance à la structure d'enchaînement, et il est le dernier qui est recommandé ici. En d'autres termes, l'indice géométrique ne retient pas des termes qui pourraient entraîner des biais à long terme.

Il apparaît également que dans la mesure du possible, il est possible d'utiliser les connaissances a priori afin d'améliorer l'indice. Les ajustements empiriques décrits à la section 3.1 peuvent être utiles, à condition qu'ils soient bien fondés. Si l'on envisage de les utiliser, il est essentiel que les connaissances a priori qui mènent à leur application soient surveillées et que leur existence continue soit vérifiée.

REMERCIEMENTS

Je tiens ici à remercier M. George Sampson pour son aide et tous les critiques et le personnel de rédaction pour leurs commentaires constructifs.

BIBLIOGRAPHIE

- DOLSON, D.D. (1982). Rent status survey: Analysis. Rapport technique, Statistique Canada.
- KOSARY, C.L., BRANSCOME, J.M. et SOMMERS, J.P. (1982). Evaluating alternatives to the rent estimator. Rapport technique, Bureau of Labor Statistics.
- KOVAR, J. (1984). Note on calculating the rent index. Rapport technique, Statistique Canada.
- SZULC, B. (1983). Linking price index numbers. Rapport technique, Statistique Canada.

Tableau 2

Erreurs quadratiques moyennes de cinq indices dans huit villes

Ville	Interpolé	Géométrique	Marginal	Arithmétique	Annuel					
Halifax	30*	(3)	19*	(2)	12*	(1)	48	(4)	74	(5)
Montréal	48*	(3)	24*	(2)	9*	(1)	160	(5)	82	(4)
Ottawa	17	(3)	12	(2)	8	(1)	22	(4)	95	(5)
Toronto	36	(4)	27	(3)	20	(2)	29	(5)	13	(1)
Winnipeg	27*	(3)	17*	(2)	10*	(1)	66	(5)	41	(4)
Edmonton	46	(1)	64	(4)	88	(5)	55	(3)	50	(2)
Calgary	56	(2)	81	(4)	121	(5)	64	(3)	46	(1)
Vancouver	70	(5)	53	(2)	39	(1)	64	(4)	60	(3)

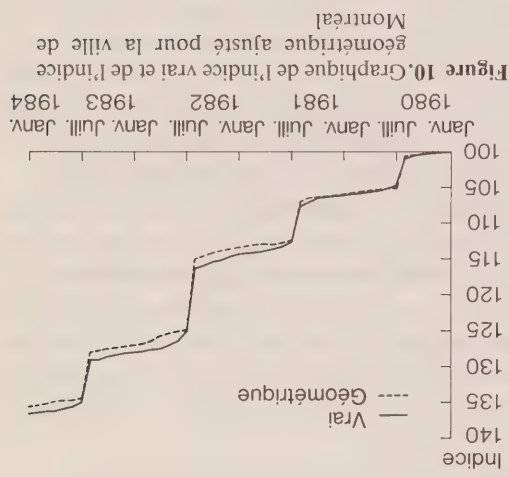
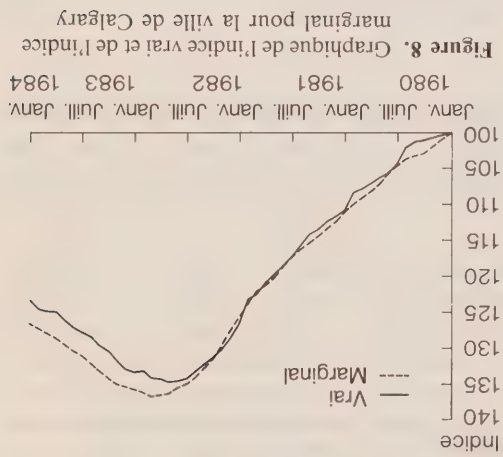
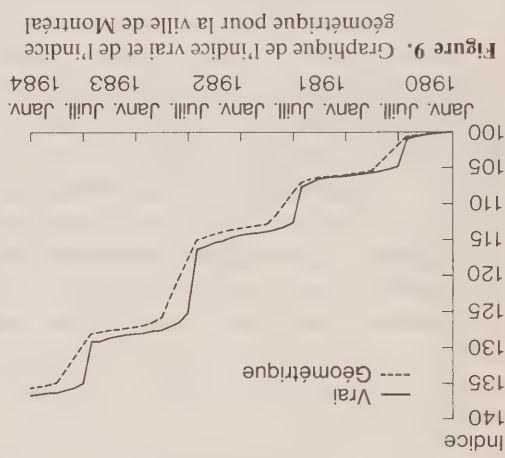
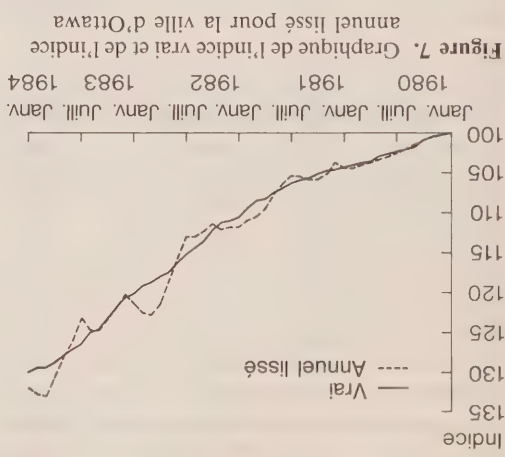
Note: 1. Les chiffres entre parenthèses donnent le classement à l'intérieur des villes
2. Les chiffres avec un astérisque sont les résultats des indices corrigés de la section 3.1

de la section 3.2 lisse effectivement l'indice, les résultats sont moins que satisfaisants, comme on peut le voir à la figure 7 (comparativement à la figure 5). On pourrait peut-être utiliser un plus grand nombre de points pour l'extrapolation, mais le problème des décalages serait encore plus important. La figure 8 montre de plus de quelle façon des variations inattendues et brusques dans les tendances générales sont reproduites avec un décalage. Cependant, les augmentations attendues de l'indice (comme en juillet à Montréal, figure 9) peuvent être ajustées de façon satisfaisante par la procédure de la section 3.1 (figure 10).

Les erreurs quadratiques moyennes des cinq indices par rapport à l'indice vrai ont été calculées pour chaque ville (tableau 2). Les trois indices basés sur l'interpolation (indices interpolé, géométrique et marginal) ont été ajustés pour les villes de Montréal, d'Halifax et de Winnipeg. Le tableau 2 présente également le classement (en ordre croissant) des erreurs quadratiques moyennes des cinq indices dans chaque ville. Les indices arithmétique et annuel ont tendance à avoir les plus mauvais résultats. Les trois indices basés sur l'interpolation donnent des résultats à peu près semblables. En général, dans les villes où l'indice augmente de façon continue, les résultats de ces trois indices se détériorent dans l'ordre: marginal, géométrique, interpolé. Cet ordre est inversé dans les villes qui enregistrent une forte chute de l'indice. Il est peu probable, cependant, que l'on puisse modifier les stratégies à partir du comportement observé seulement.

5. RÉSUMÉ

Les observations théoriques et empiriques permettent de croire que l'indice annuel est trop erratique dans les villes dont la taille de l'échantillon n'est pas assez grande. Le lissage, tout au moins celui décrit, s'est révélé inutile. Pour cette raison, l'indice annuel ne devrait être réservé qu'aux rares cas où la taille de l'échantillon le justifie. Par contre, l'indice annuel pourrait servir en rapport avec un des indices de quatre mois plus stables pour donner une estimation composite analogue à celle proposée par Kosary, Branscome et Sommers (1982). Cependant, des observations empiriques seraient nécessaires pour déterminer les poids à attribuer dans la mise en moyenne des deux indices.



éaux à 100 pour janvier 1980. L'ajustement empirique a été testé avec les données de Montréal, Halifax et Winnipeg pour les mois de juillet, janvier et octobre respectivement. Bien que les résultats de toutes les combinaisons possibles de villes et d'indices soient trop nombreux pour être inclus ici, on peut les obtenir auprès de l'auteur. Les paragraphes suivants exposeront quelques faits saillants. Bien que non exhaustifs, on espère qu'ils seront représentatifs et indicatifs de la situation réelle.

Comme on peut le voir aux figures 1-5, les cinq indices retracent l'indice vrai raisonnablement bien, même dans le cas d'un échantillon de petite taille comme pour la ville d'Ottawa. Comme on pouvait s'y attendre, les quatre premiers indices font ressortir quelques décalages, ces derniers étant plus marqués dans le cas de l'indice arithmétique et interpolé. (À noter que le problème des décalages pourrait se trouver aggravé en générant une population avec des prix augmentant de façon exponentielle). L'indice annuel est assez erratique, ce qui n'est nullement surprenant. Dans le cas des villes avec de grands échantillons (Toronto par exemple), l'indice annuel donne de bons résultats (voir figure 6). Bien que l'ajustement de lissage

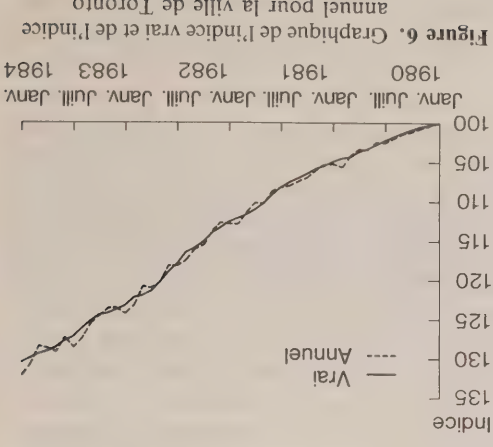
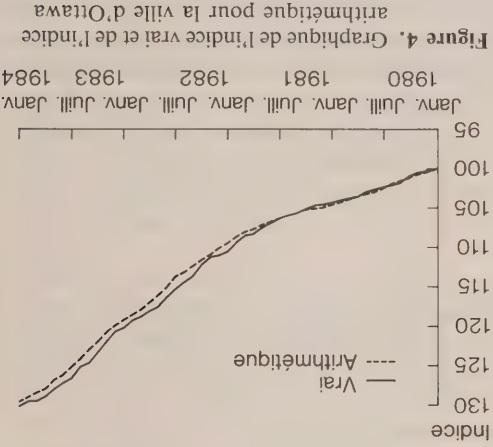
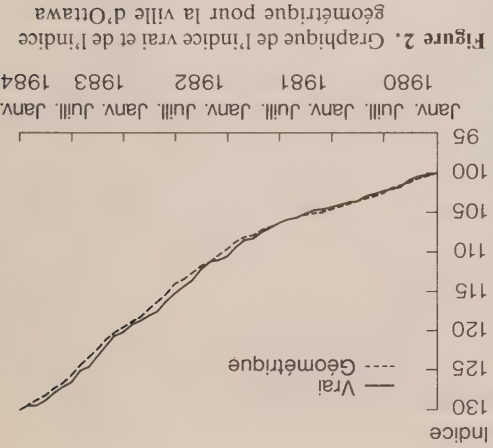
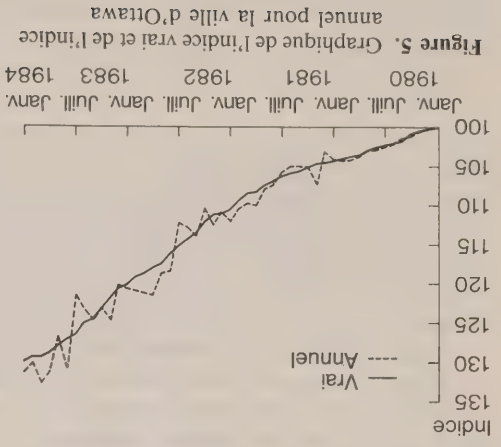
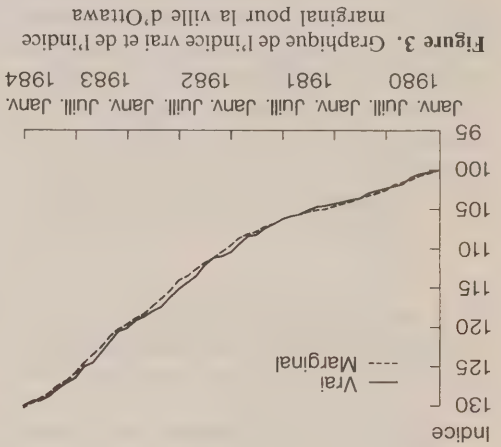
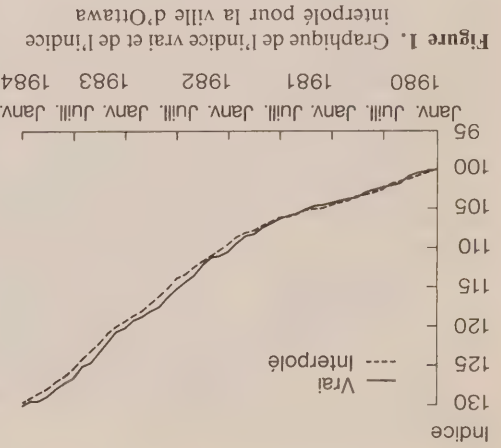


Tableau 1
Tailles moyennes d'échantillon (unités distinctes) et indice à 8401 pour huit villes d'après la population simulée

Ville	Taille moyenne de l'échantillon mensuel	Indice à 8401 (8001 = 100)
Halifax	51	144.3
Montréal	268	136.6
Ottawa	35	130.0
Toronto	170	130.4
Winnipeg	105	132.0
Edmonton	112	125.2
Calgary	97	123.5
Vancouver	105	130.5

de données. On a retenu huit villes à cette fin. Certaines sont importantes, d'autres sont petites, d'autres encore ont des bonds périodiques de leurs indices, mais toutes sont des villes avec un indice des prix et ont suffisamment de données sur les loyers disponibles. De plus, bien que certains indices de ces villes augmentent de façon monotone, d'autres augmentent et diminuent.

Seuls les loyers initiaux de toutes les unités (ceux recueillis lorsque l'unité rentre dans l'échantillon) de la base de données des loyers de l'IPC pour les années 1979 à 1984 inclusive-ment pour les huit villes retenues ont été extraits. Pour chaque unité on a ensuite simulé douze mois supplémentaires de données à partir des paramètres observés. Cette approche est sur le plan opérationnel plus simple que celle qui consiste à simuler sept mois de données en plus des six existants. Plus précisément, pour chaque unité, on a d'abord décidé si une variation de loyers allait se produire au cours des douze prochains mois. On a posé la probabilité de cet événement comme étant égale à la probabilité observée d'une variation de loyer dans cette ville particulière et pour l'année en question. Alors, en supposant qu'un changement devait se produire, on a sélectionné le mois approprié de façon à le rendre proportionnel à l'incidence observée des variations de loyers, la encore pour la ville et le mois en cause. On a supposé que le montant réel de la variation du loyer suit une distribution normale avec une moyenne et une variance fixes. On a obtenu des estimations robustes de ces deux paramètres à partir des données existantes pour chaque ville et chaque mois. Toute la programmation a été faite en SAS (Statistical Analysis System). Les nombres aléatoires ont été générés à l'aide des programmes RANUNI et RANNOR. La population obtenue comprend huit villes et quatre années de données entièrement renouvelées, c'est-à-dire en laissant de côté les mois de départ. Le tableau 1 présente les tailles de l'échantillon mensuel moyen et la valeur de l'indice simulé pour janvier 1984 (janvier 1980 = 100) pour chaque ville. Les indices, calculés pour chacune des villes, ressemblent de très près à ceux observés à l'origine. Dans les comparaisons qui suivent, on a pris les indices de population simulé comme étant les points de référence vrais à reproduire.

4.2 Comparaison des indices

Pour le calcul des indices, on a supposé que sur les treize observations disponibles pour chaque unité, seulement celles pour les mois 1, 5, 9 et 13 ont été en fait recueillies. Tous les calculs ont été ensuite faits à partir de ces sous-ensembles (4/13). Les cinq indices décrits ci-dessus ont été calculés pour chaque ville et comparés à l'indice vrai. Tous les indices sont

3.1 Ajustements empiriques

On sait, par exemple, que la plupart des loyers à Montréal changent en juillet. Les quatre premiers indices examinés à la section précédente répartiraient cette variation de juillet sur les mois de juillet, août, septembre et octobre. On pourrait cependant ajuster l'indice de juillet afin de traduire une variation plus importante et le corriger au cours des trois mois suivants. Plus précisément, on pourrait multiplier l'indice par r^* au cours du mois de référence et ensuite par $(r^*)^{-1/3}$ au cours de chacun des trois mois suivants. Comme tous les indices proposés sont des indices chaînés, le troisième mois après le mois de référence les quatre multiplicateurs se compenseraient mutuellement, ce qui ne laisserait aucun biais. Quant au choix de r^* , il dépendra d'observations empiriques continues dans chaque ville.

Il est à noter que ces ajustements ne doivent être effectués que rarement et avec beaucoup de soin. Il est essentiel que la situation particulière soit suivie de près car il n'est pas rare que de telles aberrations disparaissent brusquement.

3.2 Lissage

Comme dernière tentative de récupérer un indice erratique en dents de scie, on peut envisager de le lisser. Comme pour les ajustements ci-dessus, le lissage ne doit être envisagé que dans des cas rares et extrêmes seulement, lorsqu'il n'y a pas d'autres possibilités. La procédure de lissage que nous envisageons ici fait intervenir la mise en moyenne de l'indice au temps m par une extrapolation linéaire au temps m de l'indice lissé entre les temps $m - 1$ et $m + 2$. Un choix possible de l'indice lissé au temps m , S_m , est par conséquent donné par la formule

$$S_m = I_m/2 + (2S_{m-1} - S_{m-2})/2$$

$$= S_{m-1} + (I_m - S_{m-2})/2. \quad (3.1)$$

Comme la procédure de lissage projette en fait les données passées dans l'avenir, l'indice lissé va prolonger les tendances passées et introduira par conséquent quelques décalages. De plus, la méthode est récurrente et pourrait donc se traduire également par des biais indésirables. On pourrait envisager d'autres méthodes de lissage, bien que l'utilité de lisser un indice affecté de décalages graves soit douteuse.

4. ETUDE EMPIRIQUE

L'étude exposée dans les paragraphes suivants a été entreprise afin de tester la performance dans le temps des indices et des ajustements proposés. L'étude fournit des renseignements quantitatifs sur la capacité des indices à suivre l'indice vrai de façon précise. Elle appuie les observations les plus heuristiques faites plus haut et confirme les observations théoriques.

4.1 La population

La population des logements loués retenue dans cette étude a été définie de façon à reproduire la situation réelle le mieux possible. À cette fin, on a choisi les villes, leur importance et la taille de leurs échantillons de façon à les faire correspondre à ceux utilisés pour la composante des loyers de l'IPC. Comme toutes les données réelles sur les loyers sont disponibles pour des périodes de six mois seulement, il a fallu simuler les treize mois nécessaires

En d'autres termes, l'indice est un rapport de deux moyennes arithmétiques. De même que pour l'indice géométrique, l'indice arithmétique ne dépend que de huit mois de données seulement et il est par conséquent indépendant des mouvements entre le temps 0 et le temps $m-4$. Comme on l'a mentionné plus haut, il est également affecté de décalages de un à trois mois, et par conséquent amortit des variations soudaines.

2.5 L'indice annuel

L'expérience semble suggérer que la plupart des unités changent le loyer une fois par an. On pourrait par conséquent faire valoir que les variations annuelles sont plus stables que les variations mensuelles, puisque la distribution des variations mensuelles individuelles va nécessairement exhuber des sommets, l'un autour de la variation relative annuelle et l'autre à 1. Le régime de renouvellement du loyer pilote proposé (Kovar 1984) garantit qu'une variation annuelle soit estimable chaque mois, en d'autres termes, que r_{m-12}^m soit disponible. Afin de calculer l'indice annuel sur une base mensuelle, nous constatons que pour tout indice en-

$$(2.21) \quad I_m = r_{m-1}^m I_{m-1}$$

et

$$(2.22) \quad I_m / I_{m-12} = r_{m-12}^m.$$

À partir de ces relations nous obtenons une expression pour une variation relative mensuelle r_{m-1}^m comme (3.1).

$$(2.23) \quad r_{m-1}^m = r_{m-12}^m I_{m-12} / I_{m-1}.$$

Ces variations peuvent ensuite être enchaînées comme ci-dessus afin de donner un indice. Comme une telle relation est récursive, nous avons besoin de douze mois d'indices pour pouvoir "démarrer". Une possibilité serait de définir l'indice pour les douze premiers mois, par analogie avec l'indice géométrique, par

$$(2.24) \quad I_k = (r_{k-12}^k)^{k/12}, \quad k = 1, 2, \dots, 12.$$

Ainsi défini, l'indice annuel est indépendant des variations intermédiaires. Par contre, il sera en dents de scie, à moins que les tailles des échantillons mensuels soient importantes. Ceci s'explique par le fait que les estimations mensuelles consécutives sont totalement indépendantes. De plus, il faut constater que le problème des décalages sera au moins aussi sérieux que pour les indices exposés plus tôt.

3. AJUSTEMENTS

Dans cette section, on examine deux procédures d'ajustement pour les indices ci-dessus. D'abord, lorsque les quatre premiers indices sont affectés de décalages de un à trois mois, ils vont adoucir les sommets brusques réels. À partir des données antérieures, on a constaté que les indices des loyers présentent de brusques augmentations dans certaines villes avec une certaine régularité. Afin de "corriger" l'indice lissé, un ajustement empirique est proposé. Par contre, en raison de la nature erratique de l'indice annuel, on proposera également un ajustement pour le lissage.

Tout comme dans le cas de l'indice interpolé, l'indice marginal ne sera indépendant des observations intermédiaires que seulement en vertu de la condition restrictive selon laquelle c'est le modèle d'interpolation qui sera suivi. Dans ce cas, analogue à (2.9), le modèle s'écrit sous la forme

$$(2.18) \quad \frac{1}{I} = \frac{x_m}{1} + md.$$

Cependant, dans la plupart des situations réelles, l'indice marginal chaîne dépendra de toutes les données entre les temps -4 et m , et il sera par conséquent sensible aux divers biais

qui s'accumuleront.

Comme les trois indices examinés jusqu'à présent peuvent être décrits en termes d'interpolation linéaire de diverses fonctions des loyers observés, il est également possible de les comparer sur le plan théorique. On peut en fait montrer que les trois indices sont classés par ordre de grandeur, allant du plus petit au plus grand dans l'ordre de leur présentation. Autrement dit, dans une situation inflationniste, l'indice interpolé sera toujours inférieur en valeur absolue à l'indice géométrique, lequel à son tour sera toujours inférieur à l'indice marginal. L'inverse est vrai également lorsque la tendance est à la baisse, c'est-à-dire lorsque les prix diminuent. Comme un arbitre l'a mentionné, ce phénomène peut s'expliquer si l'on remarque que les augmentations relatives interpolées, géométriques et marginales sont respectivement les moyennes pondérées arithmétiques, géométriques et harmoniques des loyers indiqués à quatre mois d'intervalle. La relation standard entre ces moyennes explique le comportement des estimations en période d'inflation ou de déflation.

2.4 Indice arithmétique

L'indice arithmétique est construit à partir de l'échantillon avec renouvellement dont on dispose. En constatant que toutes les unités réapparaissent périodiquement dans l'échantillon, nous construisons l'indice en reportant simplement la valeur du loyer de chaque unité jusqu'à ce qu'une nouvelle observation soit enregistrée. De cette façon, toutes les unités du fichier ont un loyer correspondant pour le mois précédent, et la variation relative mensuelle r_{m-1}^m peut être construite rapidement. L'inconvénient évident est que les augmentations ou les diminutions de loyers ne sont pas enregistrées avant leur observation. Cependant, comme tous les changements sont en fin de compte enregistrés, l'indice se corrigera de lui-même (Kovar 1984), mais il sera affecté par un ensemble de décalages de un à trois mois. Tout comme dans le cas de l'indice géométrique, les variations réelles brusques seront amorties, mais l'indice arithmétique les reflétera en fin de compte.

Sur le plan technique nous constatons que dans le calcul de l'indice arithmétique pour un mois donné, un quart des observations du fichier rendent compte d'un mouvement de quatre mois, tandis que trois quarts des observations sont reportées pour un à trois mois et ne rendent compte d'aucune variation. En fait, au mois m nous observerons x_m et nous reportons x_{m-1} , x_{m-2} et x_{m-3} . De même, pour le mois $m-1$, nous observerons x_{m-1} et nous reportons x_{m-2} , x_{m-3} et x_{m-4} . La variation mensuelle est par conséquent donnée par la formule

$$(2.19) \quad r_{m-1}^m = \frac{x_m + x_{m-1} + x_{m-2} + x_{m-3}}{x_{m-1} + x_{m-2} + x_{m-3} + x_{m-4}}.$$

En enchaînant les variations comme en (2.2) et en supposant une nouvelle fois que les échantillons sont stationnaires, nous obtenons la formule de l'indice pour le mois m par rapport à l'année de base zéro

$$(2.20) \quad I_m = I_0 \frac{x_{m-3} + x_{m-2} + x_{m-1} + x_m}{x_{-3} + x_{-2} + x_{-1} + x_0}.$$

À titre de précision, constatons également que les variations relatives dans (2.12) peuvent être réécrites sous la forme

$$\frac{x_m}{x_m} = \left[\frac{x_{m-1}}{x_m} \right]_{\frac{1}{4}}$$

ou comme

$$x_{m-1} = (x_{m-4})^{\frac{1}{4}} (x_m)^{\frac{3}{4}}$$

ou enfin comme

$$\log(x_{m-1}) = (\frac{1}{4}) \log(x_{m-4}) + (\frac{3}{4}) \log(x_m). \tag{2.14}$$

L'indice géométrique équivalent par conséquent à un indice obtenu par estimation du loyer du mois précédent par une interpolation linéaire des logarithmes des loyers observés au temps m et au temps $m - 4$. (Voir (2.6) avec $y_m = \log x_m$.)

2.3 Indice marginal

Tout comme pour l'indice géométrique ci-dessus, nous supposons ici que les quatre augmentations nettes relatives mensuelles consécutives sont égales et additives. Plus précisément, nous pouvons écrire r_m^1 sous la forme

$$r_m^1 = 1 + i_m^1$$

où i_m^1 est l'augmentation nette relative du mois m par rapport au mois 1. Pour estimer r_{m-1}^{m-1} , nous devons par conséquent connaître i_{m-1}^{m-1} . En supposant que la valeur i_{m-4}^{m-4} soit égale à $4 i_{m-1}^{m-1}$, la variation relative r_{m-1}^{m-1} peut être estimée. Plus précisément, nous estimons i_{m-1}^{m-1} par

$$i_{m-1}^{m-1} = (\frac{1}{4}) i_{m-4}^{m-4} = (\frac{1}{4}) (r_{m-4}^{m-4} - 1) = (\frac{1}{4}) \left(\frac{x_m}{x_{m-4}} - 1 \right), \tag{2.15}$$

et r_{m-1}^{m-1} par

$$r_{m-1}^{m-1} = 1 + i_{m-1}^{m-1} = \frac{x_m + 3x_{m-4}}{4x_{m-4}}. \tag{2.16}$$

Nous constatons que $r_{m-1}^{m-1} = x_m/x_{m-1}$ et (2.16) peut s'écrire par conséquent sous la forme

$$\frac{x_m}{x_m + 3x_{m-4}} = \frac{4x_{m-4}}{x_{m-1}}$$

ou encore comme

$$\frac{1}{x_{m-1}} = (\frac{1}{4}) \frac{1}{x_{m-4}} + (\frac{3}{4}) \frac{1}{x_m}. \tag{2.17}$$

En d'autres termes, l'indice marginal correspond à celui que l'on aurait obtenu en estimant le loyer du mois précédent par une interpolation linéaire des inverses des loyers observés aux temps m et $m - 4$. (Voir (2.6) avec $y_m = x_m^{-1}$.)

instants, l'indice sera fixe pour tous les instants, sur la base de deux observations quelconques. Comme visiblement ce n'est pas le cas, on peut tout au mieux utiliser (2.8) comme une approximation pour les courtes périodes. Dans ce cas, cependant, si la relation en (2.9) n'est pas exacte, l'indice au temps m va dépendre de tous les loyers entre l'instant -4 et l'instant m . En d'autres termes, l'indice risque alors d'accumuler divers biais dans le temps. Remarquons que le même indice sera obtenu en supposant que l'augmentation de quatre mois, $x_m - x_{m-4}$, s'est produite lors de quatre étapes additives égales à $(x_m - x_{m-4})/4$. Dès lors, le loyer du mois précédent serait estimé par

$$(2.10) \quad x_{m-1} = x_m - (x_m - x_{m-4})/4,$$

qui est la même formule que (2.7) d'où l'autre nom: l'indice additif.

2.2 L'indice géométrique

Dans cette section, contrairement à la section précédente, nous allons essayer d'estimer la variation relative directement. Notons pour commencer que

$$(2.11) \quad r_{m-4}^m = \frac{x_m}{x_{m-4}} = \frac{x_m}{x_{m-1}} \frac{x_{m-1}}{x_{m-2}} \frac{x_{m-2}}{x_{m-3}} \frac{x_{m-3}}{x_{m-4}} = r_{m-1}^m r_{m-2}^{m-1} r_{m-3}^{m-2} r_{m-4}^{m-3}.$$

Nous supposons ensuite que les quatre variations relatives du côté droit de (2.11) sont égales, ou, ce qui revient au même, que le mouvement sur quatre mois s'explique par quatre mouvements égaux qui agissent de façon multiplicative (Kosary, Branscome et Sommers 1982). En vertu de cette hypothèse, la relation (2.11) peut s'écrire sous la forme

$$(2.12) \quad r_{m-1}^m = (r_{m-4}^m)^{1/4}.$$

À partir de (2.2) et de (2.3), en supposant qu'il n'y a pas de changements dans l'échantillon ou que les unités sortant de l'échantillon sont remplacées par des unités équivalentes qui y rentrent, l'indice du mois m par rapport à la période de base zéro devient

$$I_m = I_0 \times r_1^0 \times r_2^1 \times \dots \times r_{m-1}^{m-1} \\ = I_0 \times (r_1^{-3})^{1/4} \times (r_2^{-2})^{1/4} \times \dots \times (r_{m-4}^{m-4})^{1/4} \\ = I_0 \frac{(x_{m-3} x_{m-2} x_{m-1} x_m)^{1/4}}{(x_{-3} x_{-2} x_{-1} x_0)^{1/4}} \quad (2.13)$$

En d'autres termes, l'indice est un rapport de deux moyennes géométriques, d'où son nom d'indice géométrique. Nous constatons qu'à tout instant, en supposant que les panels soient stationnaires, l'indice ne dépend que de huit mois de données seulement, et est par conséquent indépendant de tout mouvement entre le temps 0 et le temps $m-4$, bien que dans la pratique les ensembles apparus contribuant à chaque r_{m-4}^m soient différents, de sorte que l'annulation n'est que théorique. Par contre, l'indice est affecté de décalages d'un à trois mois, ce qui aura tendance par conséquent à amortir les variations brusques vraies. Ces changements seront cependant pris en compte tôt ou tard, c'est-à-dire que l'indice va se corriger lui-même (Kovar 1984).

dans ce qui suit, tout en la conservant de façon implicite. Pour une discussion rigoureuse de ces hypothèses et de l'effet sur l'indice si les hypothèses sont mauvaises, le lecteur est prié de consulter Szulc (1983) et Kovar (1984).

Dans les paragraphes suivants, on présentera cinq méthodes d'estimation des variations mensuelles à partir de variations relatives sur quatre mois. Chacune sera justifiée de façon intuitive et théorique, et ses avantages et ses inconvénients seront présentés. Les trois premières méthodes sont établies sur une base théorique seulement, tandis que la quatrième tente d'utiliser le plan de renouvellement de l'enquête. Les quatre supposent que l'on dispose au moins de quatre mois de données rétrospectives. La dernière approche bénéficie d'une connaissance empirique a priori, celle de la probabilité élevée d'observer un changement de loyer par année. Les méthodes 2 et 4 ont déjà été étudiées auparavant par Kovar (1984).

2.1 Indice interpolé (indice additif)

Une façon d'estimer la variation relative r_{m-1}^m , est d'estimer le loyer du mois précédent, x_{m-1} . On peut y arriver, notamment, par une interpolation linéaire des loyers observés au temps m et au temps $m - 4$, c'est-à-dire en supposant que les loyers augmentent (diminuent) de façon linéaire dans le temps. Remarquons que cette hypothèse n'exige pas que chaque loyer individuel augmente chaque mois d'un montant fixe, mais que simplement la somme de tous les loyers le fait. En général, pour décrire brièvement l'interpolation linéaire, considérons deux mesures de la même quantité à deux instants différents, par exemple y_t et y_{t-s} . Supposons que nous voulions estimer la valeur de y à un instant donné entre les temps $t - s$ et t , par exemple au temps $t - u$ ($u < s$). En supposant que les mesures augmentent de façon linéaire dans le temps, il est possible d'estimer y_{t-u} à partir de y_t et y_{t-s} par la formule

(2.5)
$$y_{t-u} = (1 - \frac{s}{u}) y_t + \frac{s}{u} y_{t-s}$$

où, dans le cas qui nous intéresse, avec $s = 4$ et $u = 1$, par la formule

(2.6)
$$y_{t-1} = (\frac{3}{4}) y_t + (\frac{1}{4}) y_{t-4}.$$

On peut donc estimer le loyer total du mois précédent par la formule

(2.7)
$$x_{m-1} = (\frac{1}{4}) x_{m-4} + (\frac{3}{4}) x_m$$

et par conséquent, la variation mensuelle relative pour le mois m par

(2.8)
$$r_{m-1}^m = \frac{x_{m-1}}{x_m} = \frac{x_{m-4} + 3x_m}{4x_m}.$$

On obtient l'indice par l'enchaînement des variations relatives comme en (2.2) ci-dessus. En supposant que les loyers suivent le modèle d'interpolation linéaire, c'est-à-dire en supposant que nous pouvons exprimer le loyer du mois courant comme une fonction récurrente des loyers des mois précédents, à savoir sous la forme

(2.9)
$$x_m = x_{m-1} + d = x_0 + md,$$

on peut alors montrer que l'indice au temps m est donné par la formule $I_m^m = x_m/x_0$, comme on le voulait. En d'autres termes, si les données suivent le modèle en (2.9), l'indice ne subira pas de décalages temporels. Mais, naturellement, si le modèle était vrai pour tous les

de la tendance très rapidement. Par contre, afin de rester crédibles, les indices doivent être relativement stables, et il faut éviter les indices erratiques à dents de scie. La section 2 présente cinq estimateurs, les explique et les compare sur une base théorique. La section 3 examine quelques ajustements empiriques. Afin de comparer la performance de ces estimateurs dans le temps et dans l'espace, une étude de simulation faisant intervenir huit villes avec des observations couvrant une période de 48 mois a été effectuée. Les résultats de l'étude figurent à la section 4. La section 5 contient les conclusions et les recommandations.

2. ESTIMATEURS D'INDICES

On n'examinera ici que les indices appariés. Bien que les changements relatifs peuvent être obtenus facilement par la comparaison d'estimations indépendantes (non appariées) des niveaux des loyers à des instants différents, de telles estimations devront être extrêmement fiables, ce qui signifie des tailles d'échantillon trop importantes. De plus, les études antérieures révèlent que des estimateurs directs de ce genre ont tendance à être erratiques, biaisés à la hausse et en général peu pratiques sur le plan de l'usage (Szulc 1983). Dans ce qui suit, par conséquent, une estimation de la variation relative entre deux instants n'utilisera que les unités qui déclarent des loyers pour chacun de ces instants. Soit x_m le loyer total payé, au cours du mois courant m , par un certain sous-ensemble de logements dans une ville donnée. De façon plus rigoureuse:

(2.1)
$$x_m = \sum_{i \in s} x_{mi},$$

où x_{mi} dénote le loyer payé par le i -ième logement au cours du mois m . L'indice des loyers est estimé habituellement par l'enchaînement des indices relatifs d'un mois, c'est-à-dire les rapports des loyers moyens entre deux mois consécutifs dénotés par r_{m-1}^m . En d'autres mots, l'indice du mois m , I_m , par rapport à la période de base 0, est estimé par récurrence par

(2.2)
$$I_m = I_{m-1} \times r_{m-1}^m = 100 \times r_0^1 \times r_1^2 \times \dots \times r_{m-2}^{m-1} \times r_{m-1}^m.$$

où 100 est le niveau (arbitraire) de l'indice au temps 0. La seule difficulté est l'estimation des indices relatifs. En général, considérons la variation relative du loyer au cours du mois m par rapport au mois l , dénotée par r_l^m . Cette variation relative peut être estimée par

(2.3)
$$r_l^m = x_m / x_l.$$

Cependant, si l'on ne considère que les indices appariés seulement, les seules variations relatives en vertu du plan de sondage proposé sont les variations relatives sur quatre mois, en d'autres mots, ceux de la forme r_{m-4j}^m , $j = 1, 2, 3$ parce que c'est seulement dans ces cas-là qu'il y a des unités communes entre les deux mois. Ces variations relatives sont estimées par la formule

(2.4)
$$r_{m-4j}^m = x_m / x_{m-4j},$$

où l'ensemble s des logements comprend les unités qui déclarent un loyer à la fois au temps m et au temps $m-4j$. Malheureusement, ce qui est intéressant, c'est l'estimation des variations relatives mensuelles de la forme r_{m-1}^m . À l'actif, le plan de renouvellement garantit qu'une variation relative de quatre mois est disponible chaque mois. On suppose également que les unités qui sortent de l'échantillon sont remplacées par des unités équivalentes qui entrent dans celui-ci. Comme tel, l'ensemble s des logements communs en (2.1) dépend uniquement du temps m , et il n'est par conséquent pas nécessaire de le mentionner davantage

L'estimation d'un indice mensuel utilisant des données trimestrielles

JOHN G. KOVAR¹

RÉSUMÉ

On examine le problème de l'estimation des mouvements mensuels des loyers à partir de données recueillies tous les quatre mois. Cinq estimateurs composites différents de l'indice des loyers sont présentés et expliqués tant du point de vue intuitif que théorique. La communication présente et résume une étude empirique qui teste et compare les méthodes proposées. L'auteur expose des recommandations.

MOTS CLÉS: Nombres indices; échantillons avec renouvellement; estimation composite.

1. INTRODUCTION

La composante des loyers de l'indice des prix à la consommation utilise des données recueillies sur une base de renouvellement de six mois lors d'une enquête supplémentaire à l'enquête sur la population active. Comme les variations des loyers se produisent en général sur une base annuelle, la taille effective de l'échantillon du plan de sondage de l'enquête sur la population active s'en trouve réduite. De plus, des repères annuels spéciaux, obtenus par une revue de l'échantillon des logements de juin un an plus tard, révèlent que la composante des loyers peut être affectée par un biais plus ou moins grand (Dolsen 1982). Afin d'améliorer la situation, plusieurs plans de collecte des données ont été proposés afin de combiner les données mensuelles avec repères annuels d'une manière continue et courante. Une de ces méthodes, qui recueille des données tous les quatre mois, a été retenue pour une application pratique. Le plan proposé consiste en quatre ensembles de quatre groupes de renouvellement de logements loués, et chaque ensemble fait l'objet d'une enquête au cours d'un de quatre mois consécutifs, par renouvellement. Chaque mois, un groupe est enquêté pour la première fois, et les trois autres sont ceux qui ont été renouvelés quatre, huit et douze mois plus tôt respectivement. Chaque groupe sera donc enquêté quatre fois au cours d'une période de treize mois, avant le renouvellement de l'échantillon. Chaque mois, on obtient les données sur les loyers courants et les loyers appartés recueillis quatre mois plus tôt à partir de trois groupes de renouvellement exactement (le quatrième groupe est nouveau, et donc n'a pas de loyers "en retard" appartés). Il est possible de calculer les repères annuels sur base mensuelle à partir d'un groupe de renouvellement. On examine ici plusieurs méthodes d'estimation d'indices mensuels basées sur les données trimestrielles de ce genre. Lors de l'estimation des indices, il ne faut pas oublier les contraintes de la politique de publication de l'indice des prix à la consommation. En d'autres termes, il doit être à la fois possible sur le plan pratique et technique de produire les indices sur une base mensuelle pour chacune des villes. Les estimations doivent être fraîches, et ne doivent pas sortir après le milieu du mois suivant le mois de référence. De plus, aucune révision ne peut être apportée une fois que les indices ont été publiés. Bien que cela ne soit pas vraiment essentiel, il serait souhaitable que chaque estimateur proposé soit en mesure d'appréhender des variations soudaines (réelles)

¹ John G. Kovar, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11^e étage, immeuble R.H. Coats, Tunney's Pasture, Ottawa (Ontario) K1A 0T6

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 12, numéro 2, décembre 1986

TABLE DES MATIÈRES

J.G. KOVAR
L'estimation d'un indice mensuel utilisant des données trimestrielles..... 113

A. TEN CATE
Analyse de régression pour des données d'enquête avec plan de sondage
endogène..... 127

D.A. BINDER et G. LAZARUS
Analyse topologique des activités de la vie quotidienne à partir de l'Enquête
sur la santé et l'incapacité au Canada 145

G. HUOT et N. GAIT
La désaisonnalisation additive et la désaisonnalisation multiplicative en
présence de variations rapides de la tendance-cycle 157

Section spéciale - Les données manquantes dans les enquêtes*

D.W. CHAPMAN, L. BAILEY, et D. KASPRZYK
Méthodes de compensation de la non-réponse au U.S. Bureau of the Census 167

S. HINKINS et F. SCHEUREN
L'imputation par la méthode "hot deck" appliquée à un plan d'échantil-
lonnage à deux degrés 189

S. MICHAUD
Comparaison de la pondération et de l'imputation pour des données
non-échantillonnées 205

C.E. SÄRDAL
Estimation par la méthode de régression en situation de non-réponse..... 215

P.S.R.S. RAO
Estimation par le quotient dans le cas d'un sous-échantillonnage des
non-répondants 225

Remerciements 239

* L'édition de juin 1986 ne renferme que des articles présentés au Symposium sur les données manquantes dans les enquêtes. Étant donné le manque d'espace dans le numéro de juin, on inclut quelques articles du Symposium dans ce présent numéro.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics.

COMITÉ DE RÉDACTION

Président

R. Platek, *Statistique Canada*

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*

D.R. Bellhouse, *University of Western*

Ontario

L. Biggeri, *Université de Florence*

E.B. Dagum, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

G.J.C. Hole, *Statistique Canada*

Rédacteurs adjoints

J. Armstrong, *Statistique Canada*

H. Lee, *Statistique Canada*

COMITÉ DE DIRECTION

R. Platek (Président), J. Armstrong, E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception de coulant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociale, Statistique Canada, 4^e étage, Édifice Jean-Talton, Tunney's Pasture, Ottawa (Ontario), Canada KIA 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 10.00\$ par copie, 20.00\$ par année au Canada, et de 11.50\$ par copie, 23.00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada KIA 0T6. (Des prix réduits sont offerts aux membres de l'American Statistical Association, l'Association internationale de statisticiens d'enquête et la Société statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.)

TECHNIQUES D'Échantillonnage

UNE REVUE DE STATISTIQUE CANADA
DÉCEMBRE 1986

Publication autorisée par
le ministre des Approvisionnements
et Services Canada
©Ministre des Approvisionnements
et Services Canada 1987

March 1987
8-3200-501

Prix: Canada, \$10.00, \$20.00 par année
Autres pays, \$11.50, \$23.00 par année

Paiement en dollars canadiens ou l'équivalent
Catalogue 12-001, vol. 12, n° 2

ISSN 0714-0045

Ottawa

VOLUME 12, NUMÉRO 2
DÉCEMBRE 1986

UNE REVUE
DE
STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



JUN 10 1987

